世界知的所有権機関
国際事務局

PCT

許協力条約に基づいて公開された国際出願

(51) 国際特許分類7 G06F 17/30	A1	(11) 国際公開番号 WO00/43909 (43) 国際公開日 2000年7月27日(27.07.00)

(21) 国際出願番号 PCT/JP00/00203

(22) 国際出願日 2000年1月18日(18.01.00)

(30) 優先権データ
特願平11/13307 1999年1月21日(21.01.99) JP

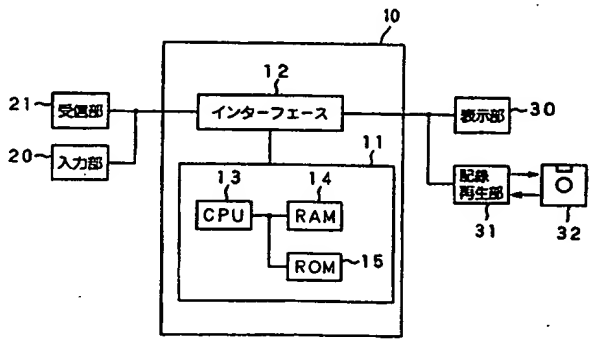
(71) 出願人（米国を除くすべての指定国について）
ソニー株式会社(SONY CORPORATION)[JP/JP]
〒141-0001 東京都品川区北品川6丁目7番35号 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
長尾 確(NAGAO, Katashi)[JP/JP]
〒141-0022 東京都品川区東五反田3丁目14番13号
株式会社 ソニーコンピュータサイエンス研究所内 Tokyo, (JP)
(74) 代理人
小池 晃, 外(KOIKE, Akira et al.)
〒105-0001 東京都港区虎ノ門二丁目6番4号
第11森ビル Tokyo, (JP)

(81) 指定国 JP, US

添付公開書類
国際調査報告書

(54)Title: METHOD AND DEVICE FOR PROCESSING DOCUMENTS AND RECORDING MEDIUM

(54)発明の名称 文書処理方法及び文書処理装置並びに記録媒体



21...RECEPTION UNIT
20...INPUT UNIT
12...INTERFACE
30...DISPLAY UNIT
31...RECORDING/REPRODUCING UNIT

(57) Abstract

A method and device for processing documents each comprised of a plurality of elements and having a tagged internal structure, wherein documents reflecting users' interests are automatically classified, by storing a plurality of documents received by a reception unit in the RAM of a control unit provided in a device body, extracting feature information indicative of features of documents according to the control of the control unit and in conformity with procedures recorded in a ROM, and classifying individual documents by classifying subject in accordance with a level of the relationship between the feature information of documents extracted by a feature information extraction unit in terms of a plurality of classifying subjects constituting a classification model and feature information for each classifying subject.

特　許　協　力　条　約

US

PC

国際調査報告

（法８条、法施行規則第40、41条）
〔ＰＣＴ１８条、ＰＣＴ規則43、44〕

出願人又は代理人 の書類記号　ＳＫ００ＰＣＴ７	今後の手続きについては、国際調査報告の送付通知様式（ＰＣＴ／ＩＳＡ／２２０）及び下記５を参照すること。	
国際出願番号 ＰＣＴ／ＪＰ００／００２０３	国際出願日 （日.月.年）　１８．０１．００	優先日 （日.月.年）　２１．０１．９９
出願人（氏名又は名称） ソニー株式会社		

国際調査機関が作成したこの国際調査報告を法施行規則第41条（ＰＣＴ１８条）の規定に従い出願人に送付する。
この写しは国際事務局にも送付される。

この国際調査報告は、全部で　4　ページである。

☐ この調査報告に引用された先行技術文献の写しも添付されている。

１．国際調査報告の基礎

ａ．言語は、下記に示す場合を除くほか、この国際出願がされたものに基づき国際調査を行った。
☐ この国際調査機関に提出された国際出願の翻訳文に基づき国際調査を行った。

ｂ．この国際出願は、ヌクレオチド又はアミノ酸配列を含んでおり、次の配列表に基づき国際調査を行った。
☐ この国際出願に含まれる書面による配列表
☐ この国際出願と共に提出されたフレキシブルディスクによる配列表
☐ 出願後に、この国際調査機関に提出された書面による配列表
☐ 出願後に、この国際調査機関に提出されたフレキシブルディスクによる配列表
☐ 出願後に提出した書面による配列表が出願時における国際出願の開示の範囲を超える事項を含まない旨の陳述書の提出があった。
☐ 書面による配列表に記載した配列とフレキシブルディスクによる配列表に記録した配列が同一である旨の陳述書の提出があった。

２．☐ 請求の範囲の一部の調査ができない（第Ⅰ欄参照）。

３．☒ 発明の単一性が欠如している（第Ⅱ欄参照）。

４．発明の名称は
☒ 出願人が提出したものを承認する。
☐ 次に示すように国際調査機関が作成した。

５．要約は
☒ 出願人が提出したものを承認する。
☐ 第Ⅲ欄に示されているように、法施行規則第47条（ＰＣＴ規則38.2(b)）の規定により国際調査機関が作成した。出願人は、この国際調査報告の発送の日から１カ月以内にこの国際調査機関に意見を提出することができる。

６．要約書とともに公表される図は、
第　1　図とする。☒ 出願人が示したとおりである。　☐ なし
☐ 出願人は図を示さなかった。
☐ 本図は発明の特徴を一層よく表している。

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ００／００２０３

第Ⅰ欄　請求の範囲の一部の調査ができないときの意見（第１ページの２の続き）

法第８条第３項（ＰＣＴ１７条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1．☐ 請求の範囲 _____ は、この国際調査機関が調査をすることを要しない対象に係るものである。
つまり、

2．☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、

3．☐ 請求の範囲 _____ は、従属請求の範囲であってＰＣＴ規則6.4(a)の第２文及び第３文の規定に従って記載されていない。

第Ⅱ欄　発明の単一性が欠如しているときの意見（第１ページの３の続き）

次に述べるようにこの国際出願に二以上の発明があるとこの国際調査機関は認めた。

請求の範囲１－８，１２，１４に共通する技術的特徴は文書分類工程であり
請求の範囲９－１１，１３，１５に共通する技術的特徴は関連度演算工程である。
上記２つの技術的特徴は相違している。

従って請求の範囲１－１５の発明の数は２である。

1．☒ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。

2．☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。

3．☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。

4．☐ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意
☐ 追加調査手数料の納付と共に出願人から異議申立てがあった。
☒ 追加調査手数料の納付と共に出願人から異議申立てがなかった。

国際調査報告　　国際出願番号　PCT／JP00／00203

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl'　G06F17／30

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl'　G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1926－1996年
日本国公開実用新案公報	1971－2000年
日本国実用新案登録公報	1996－2000年
日本国登録実用新案公報	1994－2000年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y A	JP,10-254883,A（三菱電機株式会社）,25.9月.1998(25.09.98), 請求項１（ファミリーなし）	1-5,12,14 6-8
Y A	人工知能学会誌Vol.13,No.4, 1.7月.1998(01.07.98),橋田浩一, 「GDA　意味的修飾に基づく多用途の知的コンテンツ」,pp.528-535（Journal of Japanese Society for Artificial Intelligence Vol.13,No.4,(01.07.98),Koiti Hasida,"GDA Versatile and Intelligent Contentware with Semantic annotaiton",pp.528-535)	1-5,12,14 6-11,13,15

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 30.03.00	国際調査報告の発送日　11.04.00
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員） 平井　誠　印　5L　9071 電話番号　03－3581－1101　内線　3560

様式PCT／ISA／210（第2ページ）（1998年7月）

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference Volume II, (10.08.98), Katashi Nagao, Koiti Hasida, "Automatic Text summarizatio Based on the Global document Annotation", pp. 917-921	1-15

の統語構造（syntactic structure ）を表している。タグは、エレメントの先端の前及び終端の後に対応して配置される。ここでは、エレメントの終端の後ろに配置されるタグは、記号“／”により、文書の最小単位の要素、すなわちエレメントの終端であることを示している。このエレメントは統語的構成素、すなわち句、節、及び文を示す。なお、語義＝time0は、語timeの有する複数の意味、すなわち複数の語義のうちの第０番目の意味を指している。具体的には、timeには名詞と動詞があるが、ここではtimeが名詞であることを示している。例えば、語“オレンジ”は色と果物の意味があるが、これらも語義によって区別することができる。

図２で説明したような文書のタグ付けによる内部構造は、図３に示すウインドウ１０１に示すように、その統語構造を表示することができる。このウインドウ１０１は、右半分１０３が語彙エレメントを示し、左半分１０２が文の構造を示している。このウィンドウ１０１には、タグ付けされた次に示すような文書が表示されている。この文書においても、タグによって統語構造が記述されている。次に示す文書は、「Ａ氏のＢ会が終わったＣ市で、一部の大衆紙と一般紙がその写真報道を自主規制する方針を紙面で明らかにした。」についてのタグ付けによる内部構造を示すものである。

＜文書＞＜文＞

＜形容動詞句　関係＝“場所”＞

＜名詞句＞

＜形容動詞句　場所＝“Ｃ市”＞

＜形容動詞句　関係＝“主語”＞

＜名詞句 識別子＝“Ｂ会”＞

＜形容動詞句　関係“位置”＞Ａ氏の＜／形容動詞句＞
Ｂ会
＜／名詞句＞
が
＜／形容動詞句＞
終わった
＜／形容動詞句＞
＜地名　識別子＝“Ｃ市”＞Ｃ市＜／地名＞
＜／名詞句＞
で、
＜／形容動詞句＞
＜形容動詞句　関係＝“主語”＞
＜名詞句　識別子＝新聞″統語＝“並列”＞
＜名詞句＞
＜形容動詞句＞一部の＜／形容動詞句＞
大衆紙
＜／名詞句＞と
＜名詞＞一般紙＜／名詞＞
＜／名詞句＞
が
＜／形容動詞句＞
＜形容動詞句　関係＝“目的語”＞
＜形容動詞句　関係＝“内容”　主語＝“新聞”＞
＜形容動詞句　関係＝“目的語”＞
＜名詞句＞

＜形容動詞句＞
＜名詞　共参照＝“B”＞
そ
＜／名詞＞
の
＜／形容動詞句＞
写真報道
＜／名詞句＞
を
＜／形容動詞句＞
自主規制する
＜／形容動詞句＞
方針を
＜／形容動詞句＞
＜形容動詞句　関係＝“場所”＞紙面で＜／形容動詞句＞
明らかにした。
＜／文＞＜／文書＞

この文章においては、「一部の大衆紙と一般紙」のように、統語＝“並列”は（名詞）句の並列を表す。並列の定義は、係り受け関係を共有するということである。特に何も指定がない場合は、例えば、＜名詞句　関係＝ｘ＞＜名詞＞Ａ＜／名詞＞＜名詞＞Ｂ＜／名詞＞＜／名詞句＞ はＡがＢに依存関係のあることを表す。また、関係＝ｘはこの＜名詞句＞エレメントの関係属性を表している。

続いて、タグ付けにおける、統語、意味、修辞についての相互関係を記述する関係属性について説明する。

主語、目的語、間接目的語のような文法機能、動作主、被動作者、受益者などのような主題役割及び理由、結果などのような修辞関係はこの関係属性によって表示する。関係属性は関係＝***という形で表される。本例では、比較的容易な文法機能、すなわち、主語、目的語、間接目的語のような文における当該語の機能について関係属性を記述する。

次に、本発明に係る文書処理装置の動作について、図４に示すフローチャートを参照して説明する。文書処理装置は、複数の文書について各文書の内容に関する特徴を表す特徴情報を含み、その文書の指標となるインデックスを作成する。文書分類の分類モデルに基づいて、それぞれの文書のインデックスを参照することにより文書の自動的な分類を行う。分類モデルは、文書を分類する複数の分類項目から構成され、各分類項目は特徴を表す特徴情報を有している。

最初のステップＳ１１において、文書処理装置の受信部２１は、外部から送信される複数の文書を受信する。文書処理装置は、受信部２１にて受信された複数の文書を、制御部１１の制御の下に、例えばＲＡＭ１４や記録再生部３１に記録する。文書は、図２に示したように、複数の要素すなわちエレメントからツリー状に構造化されたタグ付けによる内部構造を有している。

ステップＳ１２において、ユーザは、文書処理装置の表示部３０に表示される文書を閲覧する。文書処理装置の制御部１１は、ユーザの希望に応じて、記憶する複数の文書から表示部３０に文書を表示するように表示部３０を制御する。文書処理装置の表示部３０に表示する文書は、ユーザが入力部２０に複数の文書のうちから所望の文書の選択を入力することにより選択される。表示部３０には、

ユーザにより選択された文書の一部又は全部の内容が、その領域の大きさを変更可能なウィンドウにより表示される。なお、ユーザが文書閲覧を行うステップＳ１２は、ユーザの必要に応じて設けられる。また、図４においてステップＳ１２が平行四辺形で表されているのは、ユーザが操作することに対応している。ステップＳ１３においても同様である。

ここで、表示部３０における表示の具体例について説明する。この例は、ユーザが自由に文書を分類する分類項目であるカテゴリを設定、変更できるようにしたものである。この例においては、ユーザが設定したカテゴリに応じて文書の自動分類が行われる。

このようなグラフィックユーザインターフェース（graphic user interface; GUI）の具体例は、図５に示すようになる。このＧＵＩ画面において、操作ボタン３０２、“他のトピックス”を表示する第１の分類表示部３０３、“ビジネスニュース”を表示する第２の分類表示部３０４、“政治ニュース”を表示する第３の分類表示部３０５などをの各分類項目が表示されている。“他のトピックス”は、“他のトピックス”、“ビジネスニュース”等の特定の分類項目に分類していない文書が分類される。各分類項目の表示部においては、文書のタイトルや文書の最初の部分が表示される。

このＧＵＩにおいては、操作ボタン３０２は、画面のウィンドウの状態を初期の位置に戻すポジションリセット（position reset）と、文書の内容を閲読するブラウザ（browser ）を呼び出すブラウザのボタンと、このウィンドウから脱出するエグジット（exit）のボタンとを含んでいる。上述の各分類表示部の大きさは固定的ではなく、所望の大きさに変更することができる。分類表示部のタイト

ルも自由に変更することができる。

この自動分類は、ユーザの個別の要求に応じて分類項目を決めることにより、ユーザの関心に応えたり、ユーザが文書を探すときの効率の向上を図るものである。

ステップＳ１３においては、ユーザは、ステップＳ１２において文書処理装置の表示部３０にて閲覧した複数の文書について、この複数の文書を分類する分類項目、いわゆるカテゴリを作成し、この分類項目に上記複数の文書を分類する。文書処理装置においては、文書を分類する分類項目の設定は、分類項目の数に対応して分割された領域を有するウインドウついて、所望の分類項目を追加し、あるいは変更や削除をすることにより行われる。複数の文書の分類項目への分類は、文書の一部や分類項目のタイトルが表示され、このタイトルに対応する領域が設けられたウインドウにおいて、例えば画面上に表示されたアイコンをカーソルに連動するマウスをクリックしてドラッグすることにより、文書を所望の領域に移動して各文書を分類する。このような、文書の分類作成及び分類操作は、ユーザが表示部３０の表示を参照しながら入力部２０に入力することにより行われる。作成された分類項目及び分類操作の結果は、制御部１１の制御の下にＲＡＭ１４に記録される。なお、文書の分類項目の作成及び文書の分類の操作の詳細については後述する。

ステップＳ１４においては、文書処理装置の制御部１１は、ステップＳ１３において行われた分類項目の作成と、この分類項目に応じた分類操作に基づいて、分類モデルの作成を実行する。文書処理装置は、ＲＡＭ１４に記録されたステップＳ１３における分類項目及び分類操作の結果を読み出す。文書処理装置の制御部１１は、こ

の結果に基づいて、各分類項目に分類された上記複数の文書について、各分類項目に特徴的な固有名詞、固有名詞以外の語義、分類された文書のアドレスを集めて分類モデルを生成する。ここで、固有名詞以外の場合に語そのものではなく語義を用いるのは、同じ語でも複数の意味を有することがあるからである。文書処理装置の制御部１１は、このように作成した分類モデルを例えばＲＡＭ１４に記憶する。なお、分類モデルの作成の詳細については後述する。

以上の一連の行程により、文書を分類する基準となる分類モデルが作成された。文書処理装置は、この分類モデルを基準として、文書を自動的に分類することができる。文書処理装置が行う、新たに受信した文書の自動的な分類の動作について、図６を参照して説明する。

文書処置装置において、外部から例えば通信回線を介して受信部２１に新たな文書が送信されると、文書処理装置はこの文書を受信する。文書処理装置における文書の受信の動作については、上述したステップＳ１１で詳しく述べたので、ここでは説明を省略する。受信した文書は、例えばＲＡＭ１４や記録再生部３１に記録される。

ステップＳ２２においては、文書処理装置の制御部１１は、例えばＲＡＭ１４や記録再生部３１に記録されたステップＳ２１で受信した文書を読み出す。制御部１１は、この新たな文書から各文書の特徴を表す語を抽出することによりその文書の指標、すなわちインデックスの作成を行う。そして、制御部１１は、各文書についてのインデックスを例えばＲＡＭ１４に記録する。このインデックス作成の詳細については後述する。

ステップＳ２３においては、文書処理装置の制御部１１は、分類

モデルに基づいて、インデックスを附された各文書を上述のステップＳ１３において作成した複数の分類項目の一つに分類する。そして、制御部１１は、分類の結果を例えばＲＡＭ１４に記録する。このような文書の自動分類の詳細についてはさらに後述する。

ステップＳ２４においては、文書処理装置の制御部１１は、例えばステップＳ２３で分類され、ＲＡＭ１４に記録された新たな文書の自動分類の結果に基づいて分類モデルを更新する。制御部１１は、更新した分類モデルを例えばＲＡＭ１４に記録する。

上述したタグ付けによる内部構造を有する文書は、文書処理装置の受信部２１に外部から送信される。この文書は、例えばデジタル符号化されたいわゆる電子文書である。文書処理装置は、このような文書を例えばＲＡＭ１４や記録再生部３１に記録する。ユーザは、文書処理装置の記録する複数の文書から、任意の文書を表示部３０に表示して閲読することができる。

表示部３０における文書の表示は、大きさを変更することができるウィンドウ上に表示することができる。また、文書の表示と共に、又は文書の表示に代えて要約を表示することができる。さらに、複数の文書をウィンドウにより並べて表示したり、複数のウィンドウを重ねて表示することができる。

文書処理装置の制御部１１は、このように表示部３０に表示された文書について、ユーザの入力にしたがい各種の処理を実行する。ユーザによる文書についての入力は、表示部３０に表示されるカーソルに連動する入力部２０のマウスをクリックすることにより表示部３０における所定の領域を指定し、あるいは入力部２０のキーボードによりキーワードを入力することにより行う。

次に、上述した図４及び図６に示す処理の詳細について説明する。

まず、図６に示すＳ２２の文書の特徴を発見してインデックスを作る手順について詳細に説明する。このインデックスとは、文書の特徴を表す語を各文書について抽出して指標としたものである。文書の特徴を発見してインデックスを作成する手順は、文書処理装置の制御部１１の制御の下に、図７のフローチャートに示す一連の手順により実行される。以下の手順は、語義の関連度を算出して、この関連度に基づいてインデックスの作成を行うものである。

まず、ステップＳ３１において、制御部１１は、図６に示すステップＳ２１において受信した文書についてその文書内で活性拡散を実行し、文書内の各エレメントの活性値を拡散する。制御部１１による文書の各エレメントへの活性値の拡散処理は、後に詳細に説明する。制御部１１は、活性拡散の結果として得られた活性値を例えばＲＡＭ１４に記録する。

ステップＳ３２においては、制御部１１は、ステップＳ１１で得られた活性値に基づいて、活性値が予め設定された閾値を超えるエレメントを抽出する。制御部１１は、このように抽出したエレメントをＲＡＭ１４に記録する。

ステップＳ３３において、制御部１１は、ＲＡＭ１４からステップＳ３２で抽出したエレメントを読み出す。制御部１１は、このエレメントからすべての固有名詞を取り出してインデックスに加える。固有名詞は、語義を持たない、辞書に載っていない、などの特殊の性質を有するので、固有名詞以外の語とは別に扱う。固有名詞であるか否かは、例えば文書に付加されたタグにより識別される。例えば、図３に示したタグ付けによる内部構造においては、“Ａ氏”、

“Ｂ会”及び“Ｃ市”は固有名詞である。そして、制御部１１は、取り出した固有名詞をインデックスに加え、その結果を例えばＲＡＭ１４に記録する。

ステップＳ３４においては、制御部１１は、例えばＲＡＭ１４からステップＳ３２にて抽出したエレメントから固有名詞以外の語義を取り出してインデックスに加え、その結果をＲＡＭ１４に記録する。ここでの語義とは、語の有する複数の意味の内の選択された一つである。本実施の形態においては、語義についてもタグ付けにより記述されている。

このように、文書の特徴を発見してインデックスを作成する手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書の特徴を発見して、その特徴を配列したインデックスを作るというものである。すなわち、タグ付けによる内部構造に基づいて上記文書について活性拡散をすることにより、各語彙エレメントの活性値を拡散し、拡散後の活性値が所定の閾値より大きい語彙エレメントを抽出する。その語彙エレメントについて、固有名詞又は語義をインデックスに追加する。

上述のインデックスには、文書の特徴を表す語と共に、その文書のアドレスを含めることもできる。インデックスは、その文書を代表するような特徴を表す語を含むので、所望の文書を参照する際の指標とすることができる。このインデックスは、上述したように文書の自動的な分類に利用することができるが、これについての詳細は後述する。

ここで、インデックスの具体例を示す。

<インデックス　日付＝“AAAA/BB/CC”　時刻＝“DD:EE:FF”　文

書アドレス＝“1234”＞

＜要約＞減税規模、触れず－Ｘ首相の会見＜／要約＞

＜語　語義＝“0003”　活性値＝“140.6”＞触れず＜／語＞

＜語　語義＝“0105”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜名前　識別子＝“Ｘ”　語　語義＝“6103”　活性値＝“140.6”＞Ｘ首相＜／語＞

＜語　語義＝“5301”　活性値＝“140.6”＞求めた＜／語＞

＜語　語義＝“2350”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜語　語義＝“9582”　活性値＝“140.6”＞強調した＜／語＞

＜語　語義＝“2595”　活性値＝“140.6”＞触れる＜／語＞

＜語　語義＝“9472”　活性値＝“140.6”＞予告した＜／語＞

＜語　語義＝“4934”　活性値＝“140.6”＞触れなかった＜／語＞

＜語　語義＝“0178”　活性値＝“140.6”＞釈明した＜／語＞

＜語　語義＝“7248”　識別子＝“Ｘ”　活性値＝“140.6”＞私＜／語＞

＜語　語義＝“3684”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜語　語義＝“1824”　活性値＝“140.6”＞訴えた＜／語＞

＜語　語義＝“7289”　活性値＝“140.6”＞見せた＜／語＞

＜／インデックス＞

このインデックスにおいては、＜インデックス＞及び＜／インデックス＞はインデックスの始端及び終端を、＜日付＞及び＜時刻＞はこのインデックスが作成された日付及び時刻を、＜要約＞及び＜／

要約＞はこのインデックスの内容の要約の始端及び終端を示している。また、＜語＞及び＜／語＞は、語の始端及び終端をそれぞれ示している。語義＝“0003”は、その語義が、複数の語義のうちの第３番目であることを示している。他についても同様である。

続いて、タグ付けによる内部構造に基づいてステップＳ３１で行う、活性拡散によりエレメントの活性値を拡散する方法について説明する。この活性拡散は、後述する図１１におけるステップＳ６２においても実行される。

タグ付けによる内部構造を与えられた文書においては、活性拡散と呼ばれる処理を行うことにより、各エレメントにタグ付けによる内部構造に応じた活性値を付与することができる。活性拡散は、活性値の高いエレメントと関わりのあるエレメントにも高い活性値を与えるような処理である。この活性値は、タグ付けによる内部構造に応じて決定されるので、タグ付けによる内部構造を考慮した文書の分析に利用することができる。

活性拡散は、図８のフローチャートに示す一連の行程にしたがって、文書処理装置の制御部１１の制御の下に実行される。

最初のステップＳ４１において、文書内のエレメントの活性値の初期化が行われる。すなわち、制御部１１は、語彙エレメントを除いたすべてのエレメントと語彙エレメントに対して活性値の初期値を割り当てる。例えば、活性値の初期値として語彙エレメントを除いたすべてのエレメントに１を、語彙エレメントに０をそれぞれ割り当てればよい。また、制御部１１は、各エレメントの活性値の初期値に均一ではない値を割り当てることにより、活性拡散の結果得られた活性値の初期値の偏りを反映することができる。例えば、ユ

ーザが関心を有するエレメントに対しては、活性値の初期値を高く設定することにより、ユーザの関心を反映した活性値を得ることができる。

参照・被参照関係のエレメントを連結する参照・被参照リンクとそれ以外の通常リンクに関しては、エレメントを連結するリンクの端点の活性値である端点活性値を０に設定する。制御部１１は、このように付与した活性値の初期値をＲＡＭ１４に記録する。

エレメントとエレメントの連結は、図９に示すようになる。この図９において、文書を構成するエレメントとリンクの構造の一部として、エレメントE_i及びエレメントE_jが示されている。エレメントE_iとエレメントE_jとは、活性値e_i及びe_jをそれぞれ有し、リンクL_{ij}にて接続されている。リンクL_{ij}のエレメントE_iに接続する端点はT_{ij}、エレメントE_jに接続する端点はT_{ji}である。エレメントE_iは、リンクL_{ij}により接続されるエレメントE_jの他に、リンクL_{ik}、L_{il}及びL_{im}によって図示しないエレメントE_k、E_l及びE_mにそれぞれ接続している。エレメントE_jは、リンクL_{ji}により接続されるエレメントE_iの他に、リンクL_{jp}、L_{jq}及びL_{jr}によって図示しないエレメントE_p、E_q及びE_rにそれぞれ接続している。

ステップＳ４２においては、文書処理装置の制御部１１は、文書を構成するエレメントE_iを計数するカウンタの初期化を行う。すなわち、エレメントを計数するカウンタのカウント値ｉを１に設定する。すなわち、このカウンタは、第１番目のエレメントE_1を参照している。

ステップＳ４３において、文書処理装置の制御部１１は、カウンタが参照するエレメントについて活性値を計算するリンク処理を実

行する。このリンク処理については、さらに後述する。

ステップＳ４４において、文書処理装置の制御部１１は、文書中のすべてのエレメントについて活性値の計算が完了したか否かを判断する。制御部１１は、文書中のすべてのエレメントについて活性値の計算が完了したときには“ＹＥＳ”としてステップＳ４５に処理を進め、文書中のすべてのエレメントについて活性値の計算が完了していないときには“ＮＯ”としてステップＳ４７に処理を進める。

具体的には、制御部１１は、カウンタにて計数されている活性値の計算がなされたエレメントを参照するカウンタのカウント値ｉが、文書の含むエレメントの総数に達したか否かを判断する。制御部１１は、カウンタのカウント値ｉが文書に含まれるエレメントの総数に達したときには、すべてのエレメントが計算済みとしてステップＳ４５に処理を進め、カウンタのカウント値ｉが文書に含まれるエレメントの総数に達していないときにはすべてのエレメントについて計算が終了していないとしてステップＳ４７に処理を進める。

ステップＳ４７においては、文書処理装置の制御部１１は、カウンタのカウント値ｉを１増加させて、カウンタのカウント値をｉ＋１とする。このことにより、カウンタはｉ＋１番目のエレメント、すなわち次のエレメントを参照する。そして、処理はステップＳ４３に戻り、端点活性値の計算及びこれに続く一連の行程が、次のｉ＋１番目のエレメントについて実行される。

具体的には、制御部１１は、エレメントを計数するカウンタのカウント値ｉを１増加する。このことにより、カウンタはステップＳ４３で活性値が計算されたエレメントの次のエレメントを参照する

ことになる。

ステップＳ４５においては、文書処理装置の制御部１１は、文書に含まれるすべてのエレメントの活性値の変化分、すなわち新たに計算された活性値の元の活性値に対する変化分について、文書に含まれるすべてのエレメントについて平均値を計算する。

文書処理装置の制御部１１は、ＲＡＭ１４に記録された元の活性値と新たに計算した活性値を文書に含まれるすべてのエレメントについて読み出す。制御部１１は、新たに計算した活性値の元の活性値に対するそれぞれの変化分の総和を文書に含まれるエレメントの総数で除することにより、すべてのエレメントの活性値の変化分の平均値を計算する。制御部１１は、このように計算したすべてのエレメントの活性値の変化分の平均値をＲＡＭ１４に記録する。

ステップＳ４６においては、制御部１１は、ステップＳ４９で計算したすべてのエレメントの活性値の変化分の平均値が、予め設定された閾値以内であるか否かを判断する。制御部１１は、上記変化分が閾値以内であると“ＹＥＳ”としてこの一連の行程を終了する。制御部１１は、上記変化分が閾値以内でないときには“ＮＯ”として、ステップＳ４２にてカウンタのカウント値ｉを１に設定して文書のエレメントの活性値を計算する一連の行程を再び実行する。この一連の行程にて構成されるステップＳ４２からステップＳ４４に至るループが繰り返される毎に上記変化分は徐々に減少する。

続いて、ステップＳ４３にて実行される活性値を計算するリンク処理について、図１０に示すフローチャートを参照して説明する。

ステップＳ５１においては、文書処理装置の制御部１１は、図９に示すように、文書を構成するエレメントＥｉを計数するカウンタの

初期化を行う。すなわち、エレメントを計数するカウンタのカウント値jを1に設定する。すなわち、このカウンタは、第1番目のエレメントE_jを参照している。

ステップS52において、エレメントE_iとE_jを接続するリンクL_{ij}は、制御部11がタグを参照することによりそのリンクL_{ij}が通常リンクであるか否かを判断する。制御部11は、リンクL_{ij}について、そのリンクが語に対応する語彙エレメント、文に対応する文エレメント、段落に対応する段落エレメントなどの間の関係を示す通常リンクと、参照・被参照による係り受けの関係を示す参照リンクのいずれであるかを判断する。これは、図3の“関係”を参照することで判断することができる。制御部11は、そのリンクが通常リンクのときには“YES”としてステップS53に処理を進め、そのリンクが参照リンクのときには“NO”としてステップS54に処理を進める。

ステップS53において、通常リンクL_{ij}に対してそのリンクの端点の活性値を計算する処理が行われる。この端点活性値の計算について、図9を参照して説明する。

ここでは、ステップS52における判別により、リンクL_{ij}は通常リンクであることが明らかになっている。通常リンクL_{ij}に関して、エレメントE_iに接続する端点T_{ij}の端点活性値t_{ij}は、このリンクL_{ij}を除いたエレメントE_iに接続するすべてのリンクL_{ik}、L_{il}及びL_{im}の端点活性値t_{ik}、t_{il}及びt_{im}と、このエレメントE_iがリンクL_{ij}により接続するエレメントE_jの活性値e_jを加算し、この加算で得た値を文書に含まれるエレメントの総数で除することにより求められる。

エレメントE_iの端点T_{ij}の端点活性値は、端点T_{ij}を一端とするリンクL_{ij}が通常リンクの場合、リンクL_{ij}の他端が接続されているエレメントE_jの端点の端点活性値のうちそのリンクL_{ij}と接続されている端点T_{ji}を除いたすべての端点の端点活性値、及びそのリンクL_{ij}が接続されるエレメントE_jの活性値e_jの和を文書全体に含まれるエレメントの総数で除することにより得られる。このような手順により、活性拡散における活性値の収束が保証されることになる。

文書処理装置の制御部１１は、ＲＡＭ１４に記録されたデータから必要な端点活性値及び活性値を読み出す。制御部１１は、読み出された端点活性値及び活性値について上述のようにその通常リンクと接続された端点の端点活性値を計算する。制御部１１は、このように計算した端点活性値を例えばＲＡＭ１４に記録する。

ステップＳ５４においては、参照リンクに対して、そのリンクの端点の活性値を計算する処理が行われる。

ステップＳ５２における判別により、リンクL_{ij}は参照リンクであることが明らかになっている。通常リンクL_{ij}に関して、エレメントE_iに接続する端点T_{ij}の端点活性値t_{ij}は、このリンクL_{ij}を除いたエレメントE_iに接続するすべてのリンクL_{ik}、L_{il}及びL_{im}の端点活性値t_{ik}、t_{il}及びt_{im}と、このエレメントE_iがリンクL_{ij}により接続するエレメントE_jの活性値e_jを加算することにより求められる。

エレメントE_iの端点T_{ij}の端点活性値は、端点T_{ij}を一端とするリンクL_{ij}が参照リンクの場合、リンクL_{ij}の他端が接続されているエレメントE_jの端点の端点活性値のうちそのリンクL_{ij}と接続さ

れている端点T_{ij}を除いたすべての端点の端点活性値、及びそのリンクL_{ij}が接続されるエレメントE_jの活性値e_jの和を取ることにより得られる。

文書処理装置の制御部１１は、ＲＡＭ１４に記録されたデータから必要な端点活性値及び活性値を読み出す。制御部１１は、読み出された端点活性値及び活性値について上述のようにその参照リンクの端点活性値を計算する。制御部１１は、このように計算した端点活性値をＲＡＭ１４に記録する。

ステップＳ５３における通常リンクの処理及びステップＳ５４における参照リンクの処理は、ステップＳ４２のカウントのカウント値ｉにより参照されているエレメントE_iに接続するすべてのエレメントE_jについてのリンクL_{ij}に対して実行される。

ステップＳ５５において、文書処理装置の制御部１１は、ステップＳ５３又はステップＳ５４での計算に基づいてエレメントE_iの端点活性値を計算する。制御部１１は、この計算により得られた端点活性値をＲＡＭ１４に記録する。

ステップＳ５６においては、エレメントE_iに接続するすべてのリンクについて端点活性値t_{ij}が計算されたか否かが判別される。すべてのリンクについて端点活性値が計算されているときには“ＹＥＳ”としてステップＳ５７に進み、すべてのリンクについて端点活性値が計算されていないときには“ＮＯ”としてステップＳ５８に進む。

ステップＳ５７においては、Ｓ５６にてエレメントE_iのすべてのリンクL_{ij}について端点活性値t_{ij}が求められたことが判別されたので、エレメントE_iの活性値e_iの更新を実行する。

エレメントE_iの活性値e_iの新たな値すなわち更新値は、エレメントE_iのすべての端点の活性値の和$e_i' = e_i + \Sigma t_j'$を取ることにより求められる。ここで、“’”は、新たな値という意味である。活性値は、そのエレメントに接続するすべてのリンクについて、そのエレメントに接続する端点の端点活性値の総和となる。

文書処理装置の制御部１１は、ＲＡＭ１４に記録されたデータから必要な端点活性値t_{ij}を読み出す。制御部１１は、上述したような計算を実行し、そのエレメントE_iの活性値e_iを算出する。そして、制御部１１は、計算した新たな活性値e_iを例えばＲＡＭ１４に記録する。

次に、上述した活性値に基づいて行う語義の関連度の計算について、図１１に示すフローチャートを参照して説明する。語義の関連度の計算は、図４及び図６に示す処理を行う前にあらかじめ行う前処理であるから一度実行すればよい。

最初のステップＳ６１において、制御部１１は、電子辞書内の語の語義の説明を用い、辞書を使って語義のネットワークを作成する。すなわち、辞書における各語義の説明と、この説明中に現れる語義との参照関係から上述したような語義のタグ付けによる構造のネットワークを作成する。これは、最上位のエレメントを辞書として、図２に示したようなタグ付けによる内部構造を構成することに相当する。制御部１１は、ＲＡＭ１４に記録した語義とその説明を順に読み出してネットワークを作成する。制御部１４は、このようにして作成した語義のネットワークを例えばＲＡＭ１４や記録再生部３１に記録する。

この辞書は、例えば通信回線から受信部２１にて受信することが

でき、また、ＣＤ－ＲＯＭなどの記録媒体３２によって提供されて記録再生部３１で再生することができる。

ステップＳ６２において、ステップＳ６１で作成された語義のネットワーク上で、上述した活性拡散を行う。この活性拡散により、各語義の活性値は、上記辞書により与えられたタグ付けによる内部構造に応じて更新される。

ステップＳ６３において、ステップＳ６１で作成された語義のネットワークを構成する一つの語義s_iを選択し、ステップＳ６４においては、この語義s_iに対応する語彙エレメントE_iの活性値e_iの初期値を適当の変化させ、このときの活性値の差分Δe_iを計算する。

ステップＳ６５において、ステップＳ６４におけるエレメントE_iの活性値e_iの初期値の変化に対応する、語義s_jに対応するエレメントE_jの活性値e_jの差分Δe_jを求める。ステップＳ６６においては、ステップＳ６５で求めた差分Δe_jをステップＳ６４で求めたΔe_iで除した商$\Delta e_j / \Delta e_i$を、語義s_iの語義s_jに対する関連度とする。ある語義の活性値をステップＳ６４で変えたのに応じて、関連する語の活性値の変わることとなる。

ステップＳ６７において、語義s_iとs_jとのすべての組について関連度の演算が終了したか否かについて判断する。そして、すべての語義の組について関連度の演算が終了したときには“ＹＥＳ”として、この一連の処理を終了する。すべての語義の組について関連度の演算が終了していないときには、“ＮＯ”として、ステップＳ６３にもどり、関連度の演算が終了していない組について関連度の演算を継続する。

このように計算された関連度は、図１２に示すように、それぞれ

の語義と語義の間に定義される。この語義の表においては、関連度は正規化され、０から１までの値をとる。すなわち、この語義の表においては“コンピュータ”、“テレビ”、“ＶＴＲ”の間の相互の関連度が示されている。“コンピュータ”と“テレビ”の関連度は０．５５、“コンピュータ”と“ＶＴＲ”の関連度は０．２５、“テレビ”と“ＶＴＲ”の関連度は０．６０である。制御部１１は、このように作成した関連度を例えばＲＡＭ１４に記憶する。

ステップＳ６３からステップＳ６７のループにおいて、制御部は、必要な値を例えばＲＡＭ１４や記録再生部３１から順に読み出して上述したように関連度を計算する。制御部１１は、計算した関連度をＲＡＭ１４や記録再生部３１に記録する。

次に、上述したように算出された関連度を用いた文書分類について説明する。この関連度を利用した文書分類は、先に説明した図５のＧＵＩにおける文書分類に用いられる。

関連度による文書分類は、各分類項目の特徴を示す分類モデルを参照して関連度に基づいて行われる。分類モデルとは、各分類項目に特徴的な、固有名詞、固有名詞以外の語義、アドレスなどを含んで構成される。図１３に示す分類モデルは、各分類項目であるカテゴリに対して、固有名詞、固有名詞以外の語義、アドレスの欄を有する。この分類モデルにおいて、分類項目は“スポーツ”、“社会”、“コンピュータ”、“植物”、“美術”及び“美術”の各項目から構成されている。これらの分類項目に対応する固有名詞として、“Ａ氏”、“Ｂ社”、“Ｃ社”及び“Ｇ社”、“Ｄ種”、“Ｅ氏”、“Ｆ氏”がそれぞれ示されている。上記分類項目に対応する固有名詞以外の語義として、“野球”及び“グランド”、“労働”

及び“雇用”、“モバイル”、“桜１”及び“オレンジ１”、“桜２”及び“オレンジ２”、“桜３”がそれぞれ示されている。上記分類項目に対応するアドレスとして、“１２３４５”、“２２２２２”、“３３３３３”、“４４４４４”、“５５５５５”、“６６６６６”がそれぞれ示されている。“桜１”、“桜２”及び“桜３”は“桜”の第１の語義(１１１１１)、第２の語義(１１１１２)及び第３の語義(１１１１３)を示している。“オレンジ１”及び“オレンジ２”は、“オレンジ”の第１及び第２の語義を示している。

各分類項目の分類モデルは、タグ付けによる内部構造による活性値に基づいて抽出される。上述したように、文書処理装置の制御部１１は、ステップＳ３２において活性値が所定の閾値を超えるエレメントを抽出し、ステップＳ３３においてこのエレメントからすべての固有名詞を取り出してインデックスに加え、ステップＳ３４において固有名詞以外の語義を取り出してインデックスに加える。このように分類モデルの特徴の欄は、例えば上述の手順により生成されたインデックスを分類項目ごとにまとめたものである。

図６におけるステップＳ２３で行われる文書の自動分類は、このような分類モデルを参照して、図１４のフローチャートに示す一連の手順に従って、語義の関連度に基づいて行われる。

ステップＳ７１において、制御部１１は、分類モデルの各分類項目C_iに含まれる固有名詞の集合と、ステップＳ６２において文書から抽出されインデックスに入れられた語のうちの固有名詞の集合とについて、これらの共通集合の数を$P(C_i)$とする。そして、制御部１１は、このようにして算出した数$P(C_i)$を例えばＲＡＭ１

４に記録する。

ステップＳ７２において、制御部１１は、その文書のインデックス中の語義と各分類項目Ｃｉに含まれる語義との関連度を図１２の語義の表を参照し、語義の関連度の総和Ｒ（Ｃｉ）を演算する。制御部１１は、分類モデルにおける固有名詞以外の語について、ステップＳ６１で算出した関連度の総和Ｒ（Ｃｉ）をとる。制御部１１は、算出した関連度の総和Ｒ（Ｃｉ）をＲＡＭ１４に記録する。

ステップＳ７３において、項目Ｃｉに対する文書の関連度を、

$$Rel(C_i) = mP(C_i) + nR(C_i)$$

と定義する。ここで、係数ｍ、ｎは定数で、それぞれの値の関連度への貢献の度合いを表すパラメータである。制御部１１は、ステップＳ３３で算出した共通集合の数Ｐ（Ｃｉ）及びステップＳ６４で算出した語義の関連度の総和Ｒ（Ｃｉ）を例えばＲＡＭ１４から読み出し、上述の式に当てはめて文書の関連度Ｒｅｌ（Ｃｉ）を算出する。なお、これらの係数ｍ、ｎの値としては、例えばｍ＝１０、ｎ＝１とすることができる。そして、制御部１１は、このように算出した文書の関連度Ｒｅｌ（Ｃｉ）を例えばＲＡＭ１４に記録する。

係数ｍ及びｎの値は、統計的手法を使って推定することもできる。すなわち、制御部１１は、複数の係数ｍ及びｎの組について文書の関連度Ｒｅｌ（Ｃｉ）が与えられると、上記係数を最適化により求めることができる。

ステップＳ７４において、制御部１１は、項目Ｃｉに対する関連度Ｒｅｌ（Ｃｉ）が全項目中最大で、その関連度の値がある閾値を超えているとき、分類項目Ｃｉに文書を分類する。制御部は、複数の項目についてそれぞれ文書の関連度Ｒｅｌ（Ｃｉ）を作成し、最大の関

連度Ｒｅｌ（Ｃｉ）が閾値を超えているときには、文書を上記項目に分類Ｃｉする。最大の関連度Ｒｅｌ（Ｃｉ）が閾値を超えていないときには文書の分類は行わない。

このように、文書中に含まれる語義間の関連度の計算とそれに基づく文書の分類の手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書を処理し、この文書を複数の分類項目の内の一つの分類項目に分類する。この手順は、文書と各分類項目との関連度を算出し、算出された関連度に基づいて上記文書を分類する分類項目を決定する。

ここで、文書を分類する分類項目は、文書から抽出された固有名詞及び／又は語義を含む分類モデルによって特徴づけられる。このような分類モデルを用いて、各分類項目の分類モデルに含まれる固有名詞及び文書から抽出された固有名詞についての共通の数を算出し、各分類項目の上記分類モデルに含まれる語義に対する上記文書の関連度の総和を算出する。さらに、文書に含まれる固有名詞と関連度に基づいて抽出された固有名詞おいて重複する固有名詞の数と、語義の関連度の総和とに基づいて上記文書を分類する分類項目を決定する。この語義の関連度は、上述したようなタグ付けによる内部構造に基づいて決定される。

文書の分類は、共通する固有名詞の数及び語義の関連度の線形結合が最大となって、所定の閾値を越えるような項目に対して行われる。このような共通する固有名詞の数及び語義の関連度の線形結合の係数は、文書と分類項目の関連の大きさから、上述したように統計的に決定することができる。

次に、文書処理装置の記録再生部３１において情報が記録再生さ

れる記録媒体３２について説明する。記録媒体には、複数のエレメントからタグ付けによる内部構造を有する文書を処理する文書処理プログラムが記録されている。この記録媒体３２としては、情報の記録再生が可能な例えばフロッピーディスクが利用される。

記録媒体３２において、文書処理プログラムは、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義にそれぞれ活性値を付与する活性値付与処理と、活性値付与処理で上記語義に付与された活性値に、参照関係の組織の構造に基づいた演算を施すことにより、語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性演算処理で他の語義に付与された活性値の微分を活性値付与工程で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有する。

記録媒体３２において、参照関係組織処理で用いられる辞書の各語義にはその語義の属性を示す属性情報が付与され、上記参照関係の組織は上記属性情報に基づいて作成される。

記録媒体３２に記録される文書処理プログラムは、エレメントの最小単位である語義の間の相互の関連度を算出する関連度算出処理と、文書を分類する複数の分類項目について、分類項目の特徴をあらわす語義を含んでなる分類モデルを用い、各分類モデルの含む語義との関連度に基づいて文書を分類する文書分類処理との各処理工程を有する。

記録媒体３２に記録される文書処理プログラムの関連度算出処理は、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、上記語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義の構造に基づいた演算を施すことにより、語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性値演算処理で他の語義に付与された活性値の微分を活性値付与処理で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有する。

なお、本例において、文書へのタグ付けの方法の一例を示したが、本発明がこのタグ付けの方法に限定されないことはもちろんである。また、本例において、文書処理装置の受信部２１に外部から文書が送信されるとしたが、本発明はこれに限定されない。例えば、上記文書は、文書処理装置のＲＯＭ１３に書き込まれ記録再生部３１において記録媒体３２から読み出されてもよい。

上述した説明において、文書処理装置の表示部３０に表示された文書から所望のエレメントを選択するデバイスとしてマウスを例示したが、本発明がこれに限定されないことはいうまでもない。文書処理装置におけるエレメントの入力には、タブレット、ライトペン等の他のデバイスを利用することができる。

産業上の利用可能性

上述したような本発明を用いることにより、語義の関連度を算出することができ、この語義の関連度を利用することにより、語義の関連度に基づいて、ユーザの興味を反映した文書の自動分類のような文書処理を実行することができる。語義の関連度に基づく文書処理は、自動的に実行することができるので、文書処理の際のユーザの負担を軽減する。

請求の範囲

1．複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出工程と、

分類モデルを構成する複数の分類項目について上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて各文書を上記分類項目に分類する文書分類工程とを有する文書処理方法。

2．上記文書処理方法は、さらに複数の文書を受信する受信工程を有し、上記特徴情報抽出工程は、上記受信工程で受信した各文書についてその文書の特徴を表す特徴情報を抽出する請求の範囲第1項記載の文書処理方法。

3．上記特徴情報抽出工程は、上記文書の内部構造に基づいて各要素に重みを付与しこの重みが所定値より大きい要素を抽出する請求の範囲第1項記載の文書処理方法。

4．上記文書分類工程は、複数の文書を分類項目に分類した分類操作に基づいて作成された分類モデルを用いて文書を分類する請求の範囲第1項記載の文書処理方法。

5．上記文書分類工程は、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目の特徴情報との関連度を計算し、上記関連度が閾値を超えると上記関連度が最大となる分類項目に文書を分類する請求の範囲第1項記載の文書処理方法。

6．上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項

目の特徴情報の関連度は、上記文書の特徴情報に含まれる固有名詞と上記分類項目の特徴情報に含まれる固有名詞とに共通する固有名詞の数と、上記文書の特徴情報に含まれる固有名詞以外の語義の上記分類項目に含まれる固有名詞以外の語義に対する語義の関連度の総和との線形結合である請求の範囲第５項記載の文書処理方法。

７．上記語義の関連度は、語義の間の参照関係の構造に基づいて各語義に重みを付与し、一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度とすることにより得られたものである請求の範囲第６項記載の文書処理方法。

８．上記語義の間の参照関係は、各語義について他の語義を参照する辞書を用いて作成される請求の範囲第７項記載の文書処理方法。

９．複数の要素から構成される内部構造を有する文書を処理する文書処理方法であり、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と、

上記重み付与工程で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算工程とを有する文書処理方法。

１０．上記重み付与工程は、各語義について他の語義を参照する辞書を用いて語義の参照関係を組織した参照関係の構造に基づいて重みを付与する請求の範囲第９項記載の文書処理方法。

１１．上記辞書の各語義にはその属性を示す属性情報が付与され、上記参照関係の構造は上記属性情報に基づいて組織される請求の範囲第１０項記載の文書処理方法。

１２．複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出手段と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出手段で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類手段とを有する文書処理装置。

１３．複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与手段と、

上記重み付与手段で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算手段とを有する文書処理装置。

１４．複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、

上記文書処理プログラムは、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出処理で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類処理とを有する記録媒体。

１５．複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、

上記文書処理プログラムは、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と、

上記重み付与処理で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理とを有する記録媒体。

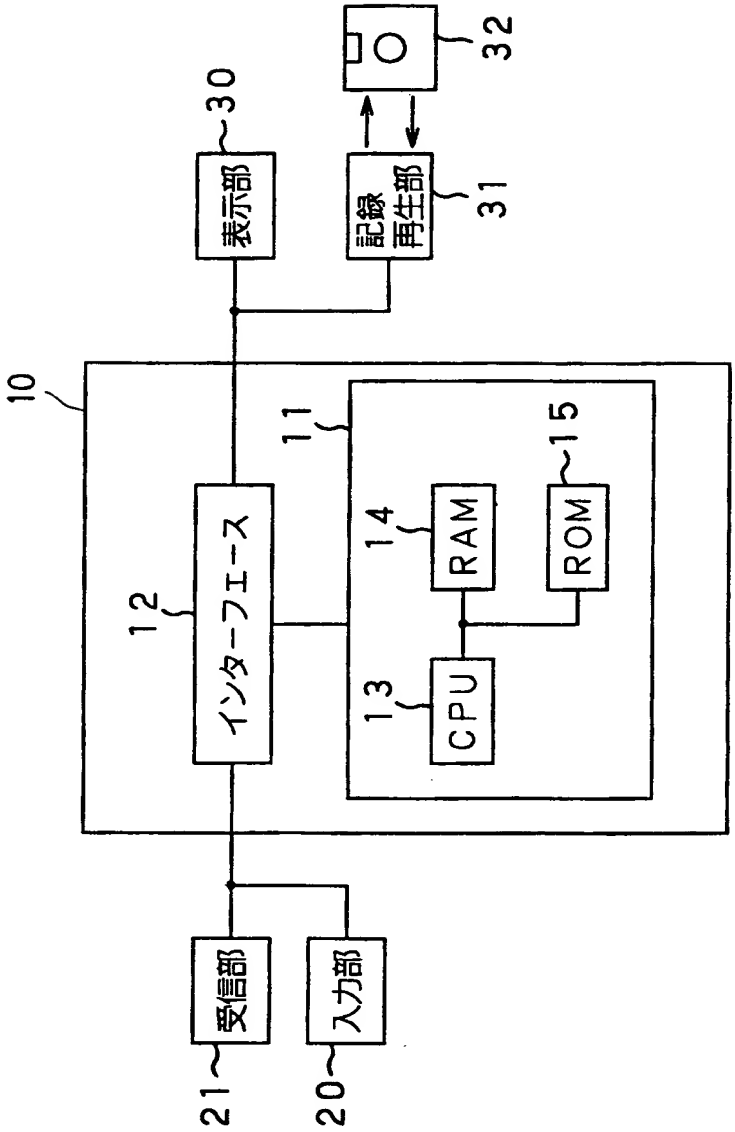

32
30
表示部
記録
再生部
31
10
11
15
12
インターフェース
14
RAM
ROM
13
CPU
受信部
入力部
21
20

図 1

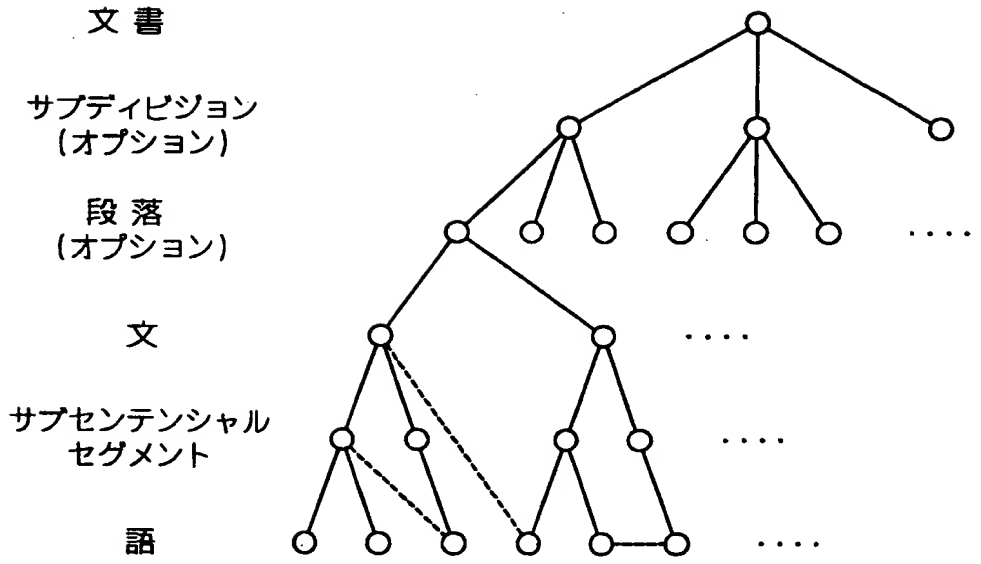


図 2

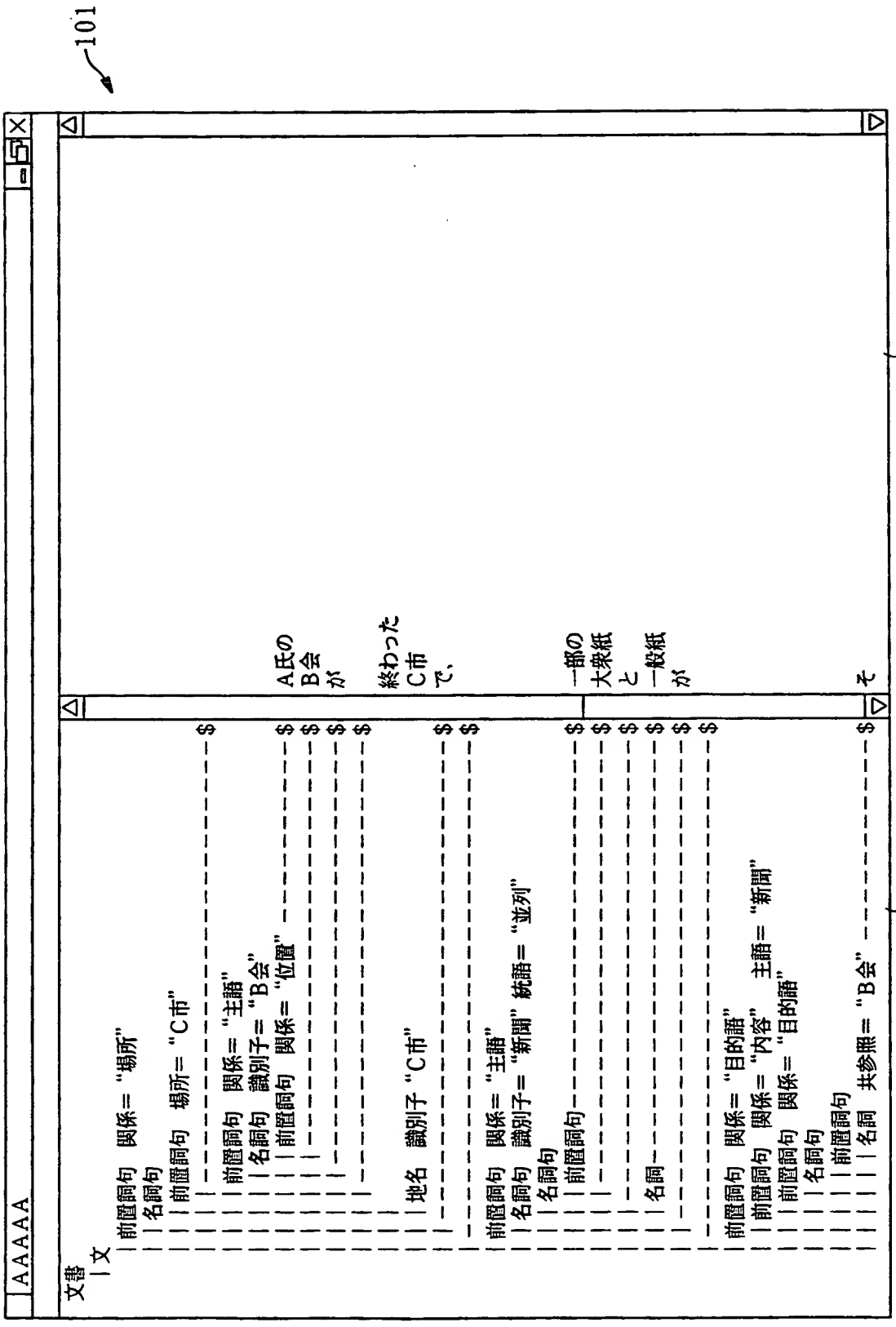

101
AAAAA
文書
文
前置詞句 関係= "場所"
名詞句
前置詞句 場所= "C市"
名詞句
前置詞句 関係= "主語"
名詞句 識別子= "B会"
前置詞句 関係= "位置"
地名 識別子 "C市"
前置詞句 関係= "主語"
名詞句 識別子= "新聞" 統語= "並列"
名詞句
前置詞句
名詞
前置詞句 関係= "目的語"
前置詞句 関係= "内容"
前置詞句 関係= "目的語"
名詞句
前置詞句
名詞 共参照= "B会"
A氏の
B会
が
終わった
C市
で、
一部の
大衆紙
と
一般紙
が
そ
102
103

図 3

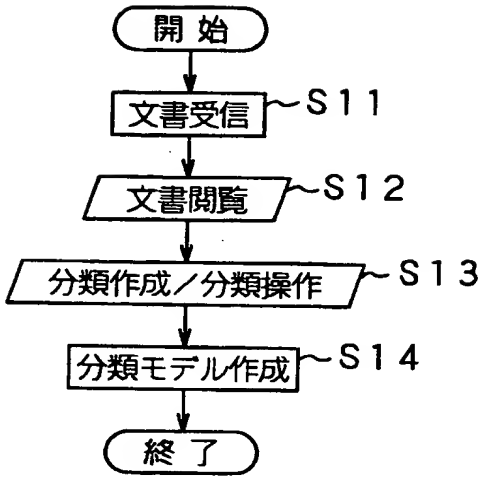

開始
文書受信
S11
文書閲覧
S12
分類作成／分類操作
S13
分類モデル作成
S14
終了

図 4

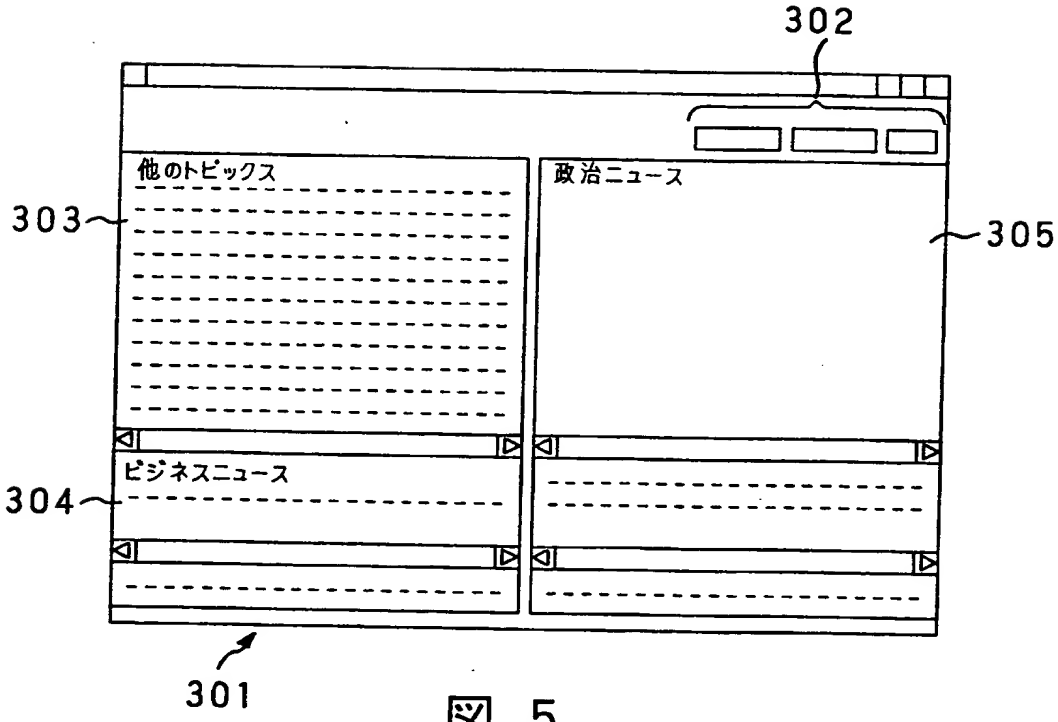

302
他のトピックス
303
政治ニュース
305
ビジネスニュース
304
301

図 5

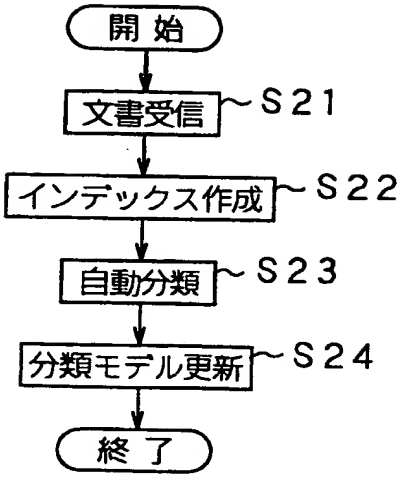

開 始
文書受信
S21
インデックス作成
S22
自動分類
S23
分類モデル更新
S24
終 了

図 6

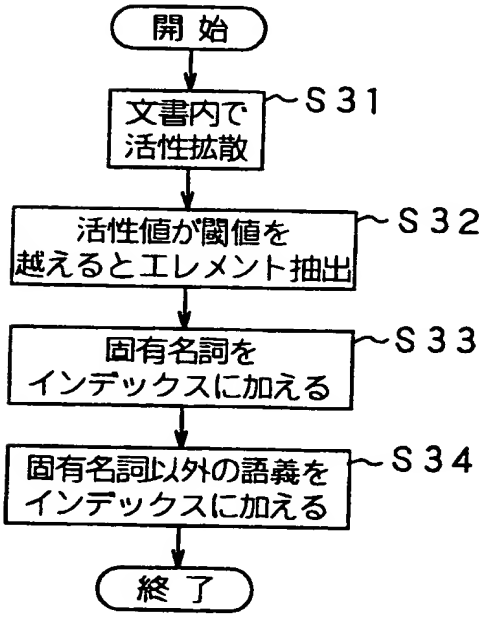

開 始
文書内で
活性拡散
S31
活性値が閾値を
越えるとエレメント抽出
S32
固有名詞を
インデックスに加える
S33
固有名詞以外の語義を
インデックスに加える
S34
終 了

図 7

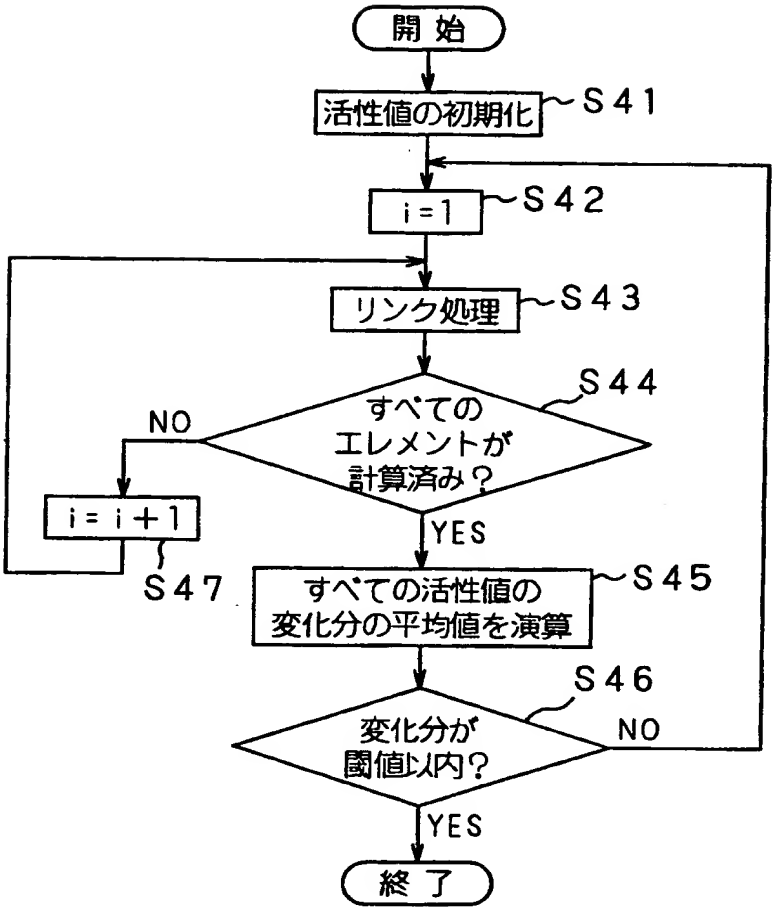

開 始
活性値の初期化
S41
i=1
S42
リンク処理
S43
S44
すべての
エレメントが
計算済み？
NO
YES
i=i+1
S47
すべての活性値の
変化分の平均値を演算
S45
S46
変化分が
閾値以内？
NO
YES
終 了

図 8

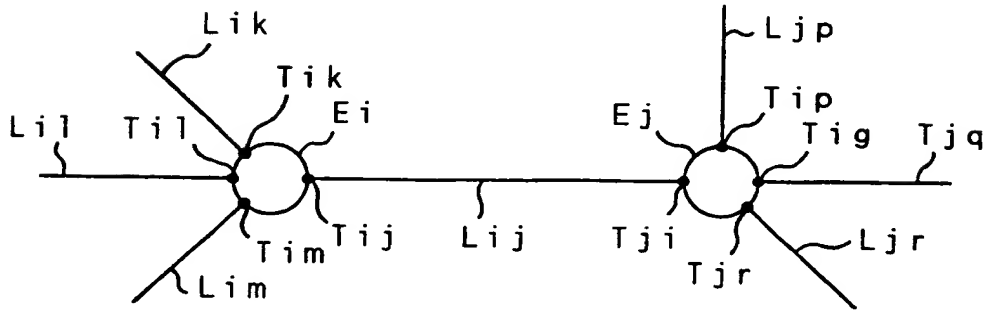

Lik
Tik
Ei
Lil
Til
Tim
Tij
Lij
Lim
Ljp
Tip
Ej
Tig
Tjq
Tji
Tjr
Ljr

図 9

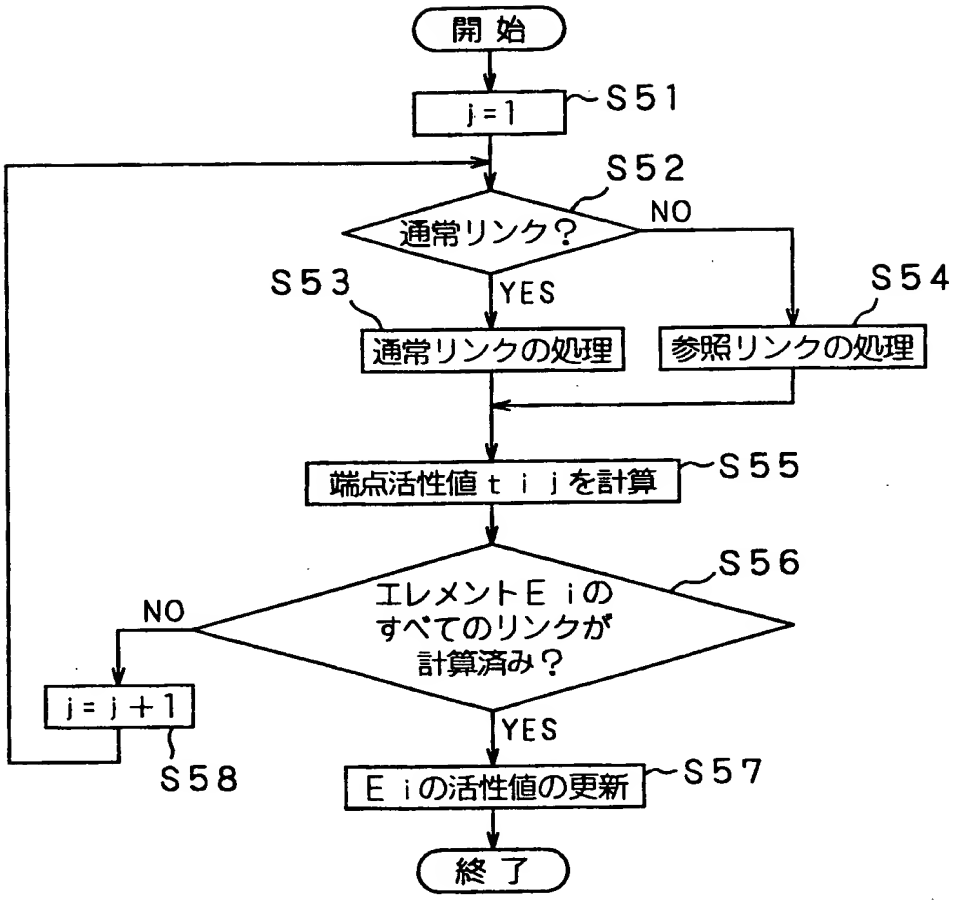

開始
j=1
S51
S52
通常リンク?
NO
YES
S53
S54
通常リンクの処理
参照リンクの処理
端点活性値ｔｉｊを計算
S55
S56
エレメントＥｉの
すべてのリンクが
計算済み?
NO
YES
j=j+1
S58
Ｅｉの活性値の更新
S57
終了

図10

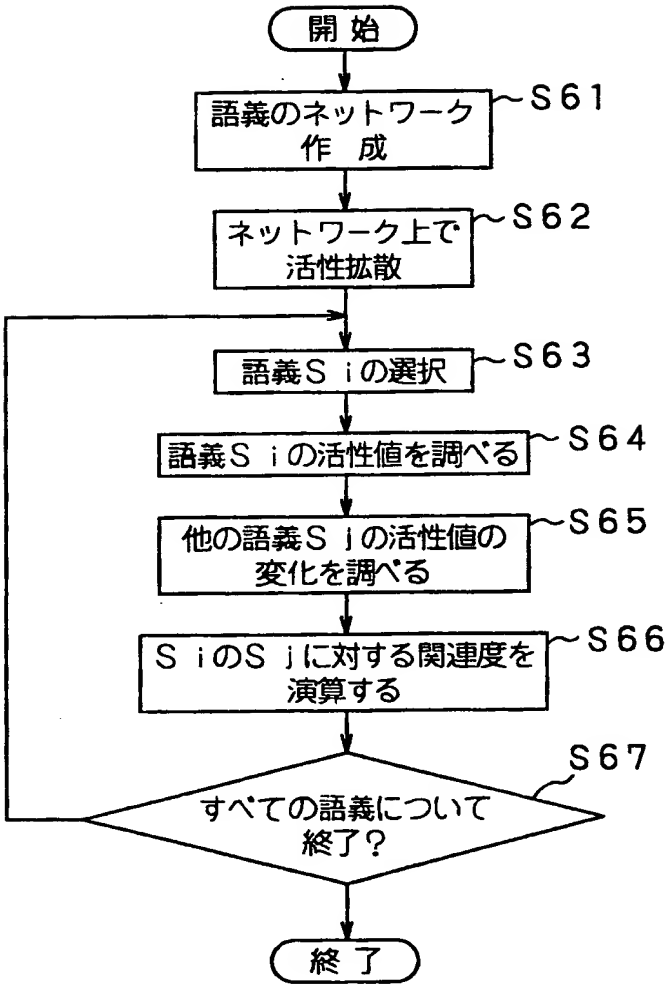


図11

	コンピュータ	テレビ	
コンピュータ		0.55	
テレビ	0.55		
VTR	0.25	0.60	

図12

分類項目	スポーツ	社 会	コンピュータ	植 物	美 術	イベント
固有名詞	A 氏	B 社	C社 G社	D 種	E 氏	F 氏
語 義	野 球 グランド	労働 雇用	モバイル	桜1 (11111) オレンジ1	桜2 (11112) オレンジ2	桜3 (11113)
文書 アドレス	12345	22222	33333	44444	55555	66666

図13

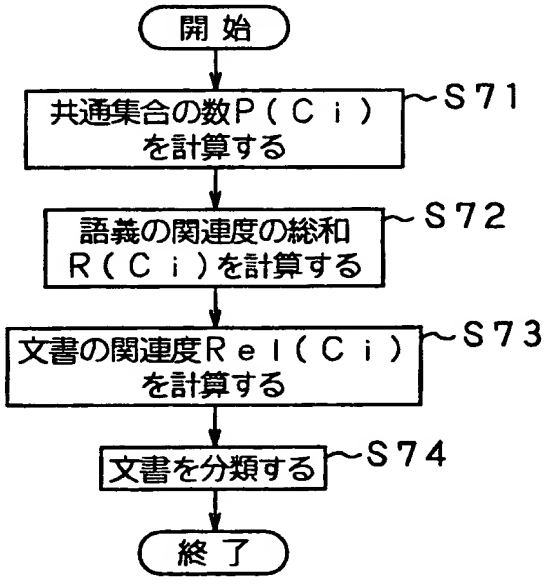


図14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP00/00203

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926-1996 Jitsuyo Shinan Toroku Koho 1996-2000
Kokai Jitsuyo Shinan Koho 1971-2000 Toroku Jitsuyo Shinan Koho 1994-2000

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP, 10-254883, A (Mitsubishi Electric Corporation), 25 September, 1998 (25.09.98), Claim 1 (Family: none)	1-5,12,14 6-8
Y A	Koichi Hashida, "GDA Versatile and Intelligent Content ware with Semantic Annotation", Journal of Japanese Society for Artificial Intelligence, Vol.13, No.4, 01 July, 1998 (01.07.98), pp.528-535	1-5,12,14 6-11,13,15
A	Katashi Nagao, Koiti Hasida, "Automatic Text summarization Based on the Global document Annotation", 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference Volume II,(10.08.98), pp.917-921	1-15

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 March, 2000 (30.03.00)	11 April, 2000 (11.04.00)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP00/00203

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

A technical feature common to Claims 1 to 8, 12 and 14 is a document classifying step, while a technical feature common to Claims 9 to 11, 13 and 15 is a relationship level computing step.
The above two technical features are quite different from each other.
Therefore, the number of inventions in Claims 1 to 15 is two.

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☒ No protest accompanied the payment of additional search fees.

国際調査報告

国際出願番号　PCT／JP00／00203

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1926－1996年
日本国公開実用新案公報	1971－2000年
日本国実用新案登録公報	1996－2000年
日本国登録実用新案公報	1994－2000年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y A	JP, 10-254883, A（三菱電機株式会社）, 25. 9月. 1998(25. 09. 98), 請求項1（ファミリーなし）	1-5, 12, 14 6-8
Y A	人工知能学会誌Vol. 13, No. 4, 1. 7月. 1998(01. 07. 98), 橋田浩一, 「GDA　意味的修飾に基づく多用途の知的コンテンツ」, pp. 528-535 (Journal of Japanese Society for Artificial Intelligence Vol. 13, No. 4, (01. 07. 98), Koiti Hasida, "GDA Versatile and Intelligent Contentware with Semantic annotaiton", pp. 528-535)	1-5, 12, 14 6-11, 13, 15

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日 30. 03. 00	国際調査報告の発送日　11.04.00
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 平井　誠　印　5L　9071 電話番号　03－3581－1101　内線　3560

様式PCT／ISA／210（第2ページ）（1998年7月）

国際調査報告

国際出願番号 PCT／JP00／00203

C（続き）. 関連すると認められる文献

引用文献のカテゴリー＊	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference Volume II, (10.08.98), Katashi Nagao, Koiti Hasida, "Automatic Text summarizatio Based on the Global document Annotation", pp. 917-921	1-15

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ００／００２０３

第Ⅰ欄　請求の範囲の一部の調査ができないときの意見（第１ページの２の続き）

法第８条第３項（ＰＣＴ１７条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1．☐ 請求の範囲 _____ は、この国際調査機関が調査をすることを要しない対象に係るものである。
つまり、

2．☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、

3．☐ 請求の範囲 _____ は、従属請求の範囲であってＰＣＴ規則6.4(a)の第２文及び第３文の規定に従って記載されていない。

第Ⅱ欄　発明の単一性が欠如しているときの意見（第１ページの３の続き）

次に述べるようにこの国際出願に二以上の発明があるとこの国際調査機関は認めた。

請求の範囲１－８，１２，１４に共通する技術的特徴は文書分類工程であり
請求の範囲９－１１，１３，１５に共通する技術的特徴は関連度演算工程である。
上記２つの技術的特徴は相違している。

従って請求の範囲１－１５の発明の数は２である。

1．☒ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。

2．☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。

3．☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。

4．☐ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意

☐ 追加調査手数料の納付と共に出願人から異議申立てがあった。

☒ 追加調査手数料の納付と共に出願人から異議申立てがなかった。

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP00/00203

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926-1996 Jitsuyo Shinan Toroku Koho 1996-2000
Kokai Jitsuyo Shinan Koho 1971-2000 Toroku Jitsuyo Shinan Koho 1994-2000

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP, 10-254883, A (Mitsubishi Electric Corporation), 25 September, 1998 (25.09.98), Claim 1 (Family: none)	1-5,12,14 6-8
Y A	Koichi Hashida, "GDA Versatile and Intelligent Content ware with Semantic Annotation", Journal of Japanese Society for Artificial Intelligence, Vol.13, No.4, 01 July, 1998 (01.07.98), pp.528-535	1-5,12,14 6-11,13,15
A	Katashi Nagao, Koiti Hasida, "Automatic Text summarization Based on the Global document Annotation", 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference Volume II,(10.08.98), pp.917-921	1-15

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 March, 2000 (30.03.00)	11 April, 2000 (11.04.00)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP00/00203

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

A technical feature common to Claims 1 to 8, 12 and 14 is a document classifying step, while a technical feature common to Claims 9 to 11, 13 and 15 is a relationship level computing step.
The above two technical features are quite different from each other.
Therefore, the number of inventions in Claims 1 to 15 is two.

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☒ No protest accompanied the payment of additional search fees.

特許協力条約に基づく国際出願願書 SK00PCT7
副本 - 印刷日時 2000年01月18日 (18.01.2000) 火曜日 16時06分28秒

0	受理官庁記入欄	
0-1	国際出願番号.	
0-2	国際出願日	
0-3	(受付印)	

PCT 18.1.00 受領印

0-4	様式-PCT/RO/101 この特許協力条約に基づく国際出願願書は、	
0-4-1	右記によって作成された。	PCT-EASY Version 2.90 (updated 15.12.1999)
0-5	申立て 出願人は、この国際出願が特許協力条約に従って処理されることを請求する。	
0-6	出願人によって指定された受理官庁	日本国特許庁 (RO/JP)
0-7	出願人又は代理人の書類記号	SK00PCT7
I	発明の名称	文書処理方法及び文書処理装置並びに記録媒体
II	出願人	
II-1	この欄に記載した者は	出願人である (applicant only)
II-2	右の指定国についての出願人である。	米国を除くすべての指定国 (all designated States except US)
II-4ja	名称	ソニー株式会社
II-4en	Name	SONY CORPORATION
II-5ja	あて名:	141-0001 日本国 東京都 品川区 北品川６丁目７番３５号
II-5en	Address:	7-35, Kitashinagawa 6-chome Shinagawa-ku, Tokyo 141-0001 Japan
II-6	国籍 (国名)	日本国 JP
II-7	住所 (国名)	日本国 JP

特許協力条約に基づく国際出願願書 SK00PCT7

副本 - 印刷日時 2000年01月18日 (18.01.2000) 火曜日 16時06分28秒

III-1	その他の出願人又は発明者	
III-1-1	この欄に記載した者は	出願人及び発明者である (applicant and inventor)
III-1-2	右の指定国についての出願人である。	米国のみ (US only)
III-1-4ja	氏名(姓名)	長尾 確
III-1-4en	Name (LAST, First)	NAGAO, Katashi
III-1-5ja	あて名:	141-0022 日本国 東京都 品川区 東五反田３丁目１４番１３号 株式会社ソニーコンピュータサイエンス研究所内
III-1-5en	Address:	c/o SONY COMPUTER SCIENCE LABORATORY INC. 14-13, Higashi-Gotanda 3-chome Shinagawa-ku, Tokyo 141-0022 Japan
III-1-6	国籍 (国名)	日本国 JP
III-1-7	住所 (国名)	日本国 JP
IV-1	代理人又は共通の代表者、通知のあて名 下記の者は国際機関において右記のごとく出願人のために行動する。	代理人 (agent)
IV-1-1ja	氏名(姓名)	小池 晃
IV-1-1en	Name (LAST, First)	KOIKE, Akira
IV-1-2ja	あて名:	105-0001 日本国 東京都 港区 虎ノ門二丁目６番４号 第１１森ビル
IV-1-2en	Address:	No.11 Mori Bldg., 6-4, Toranomon 2-chome Minato-ku, Tokyo 105-0001 Japan
IV-1-3	電話番号	03-3508-8266
IV-1-4	ファクシミリ番号	03-3508-0439
IV-2	その他の代理人	筆頭代理人と同じあて名を有する代理人 (additional agent(s) with same address as first named agent)
IV-2-1ja	氏名	田村 榮一; 伊賀 誠司
IV-2-1en	Name(s)	TAMURA, Eiichi; IGA, Seiji
V	国の指定	
V-1	広域特許 (他の種類の保護又は取扱いを求める場合には括弧内に記載する。)	--
V-2	国内特許 (他の種類の保護又は取扱いを求める場合には括弧内に記載する。)	JP US

特許協力条約に基づく国際出願願書 SK00PCT7

副本 - 印刷日時 2000年01月18日 (18.01.2000) 火曜日 16時06分28秒

V-5	指定の確認の宣言 出願人は、上記の指定に加えて、規則4.9(b)の規定に基づき、特許協力条約のもとで認められる他の全ての国の指定を行う。ただし、V-6欄に示した国の指定を除く。出願人は、これらの追加される指定が確認を条件としていること、並びに優先日から15月が経過する前にその確認がなされない指定は、この期間の経過時に、出願人によって取り下げられたものとみなされることを宣言する。		
V-6	指定の確認から除かれる国	なし (NONE)	
VI-1	先の国内出願に基づく優先権主張		
VI-1-1	先の出願日	1999年01月21日 (21.01.1999)	
VI-1-2	先の出願番号	平成11年特許願第013307号	
VI-1-3	国名	日本国 JP	
VII-1	特定された国際調査機関(ISA)	日本国特許庁 (ISA/JP)	
VIII	照合欄	用紙の枚数	添付された電子データ
VIII-1	願書	4	-
VIII-2	明細書	35	-
VIII-3	請求の範囲	4	-
VIII-4	要約	1	absk00pct7.txt
VIII-5	図面	9	-
VIII-7	合計	53	
	添付書類	添付	添付された電子データ
VIII-8	手数料計算用紙	✓	-
VIII-12	優先権証明書	優先権証明書 VI-1	-
VIII-16	PCT-EASYディスク	-	フレキシブルディスク
VIII-17	その他	納付する手数料に相当する特許印紙を貼付した書面	-
VIII-17	その他	国際事務局の口座への振込を証明する書面	-
VIII-18	要約書とともに提示する図の番号	1	
VIII-19	国際出願の使用言語名:	日本語 (Japanese)	
IX	提出者の記名押印		
IX-1	氏名(姓名)		
IX-2	権限		

受理官庁記入欄

10-1	国際出願として提出された書類の実際の受理の日	
10-2	図面 :	
10-2-1	受理された	
10-2-2	不足図面がある	

特許協力条約に基づく国際出願願書
副本 - 印刷日時 2000年01月18日 (18.01.2000) 火曜日 16時06分28秒

10-3	国際出願として提出された書類を補完する書類又は図面であってその後期間内に提出されたものの実際の受理の日（訂正日）	
10-4	特許協力条約第11条(2)に基づく必要な補完の期間内の受理の日	
10-5	出願人により特定された国際調査機関	ISA/JP
10-6	調査手数料未払いにつき、国際調査機関に調査用写しを送付していない	

国際事務局記入欄

11-1	記録原本の受理の日	

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF RECEIPT OF RECORD COPY

(PCT Rule 24.2(a))

To:

KOIKE, Akira
No.11 Mori Bldg., 6-4, Toranomon 2-chome
Minato-ku, Tokyo 105-0001
JAPON

Date of mailing (day/month/year)	IMPORTANT NOTIFICATION
02 February 2000 (02.02.00)	
Applicant's or agent's file reference	**International application No.**
SK00PCT7	PCT/JP00/00203

The applicant is hereby notified that the International Bureau has received the record copy of the international application as detailed below.

Name(s) of the applicant(s) and State(s) for which they are applicants:

SONY CORPORATION (for all designated States except US)
NAGAO, Katashi (for US)

International filing date	:	18 January 2000 (18.01.00)
Priority date(s) claimed	:	21 January 1999 (21.01.99)
Date of receipt of the record copy by the International Bureau	:	28 January 2000 (28.01.00)
List of designated Offices	:	

National :JP,US

ATTENTION

The applicant should carefully check the data appearing in this Notification. In case of any discrepancy between these data and the indications in the international application, the applicant should immediately inform the International Bureau.

In addition, the applicant's attention is drawn to the information contained in the Annex, relating to:

- [X] time limits for entry into the national phase
- [X] confirmation of precautionary designations
- [ ] requirements regarding priority documents

A copy of this Notification is being sent to the receiving Office and to the International Searching Authority.

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 740.14.35

Authorized officer:

Y. KUWAHARA

Telephone No. (41-22) 338.83.38

INFORMATION ON TIME LIMITS FOR ENTERING THE NATIONAL PHASE

The applicant is reminded that the "national phase" must be entered before each of the designated Offices indicated in the Notification of Receipt of Record Copy (Form PCT/IB/301) by paying national fees and furnishing translations, as prescribed by the applicable national laws.

The time limit for performing these procedural acts is **20 MONTHS** from the priority date or, for those designated States which the applicant elects in a demand for international preliminary examination or in a later election, **30 MONTHS** from the priority date, provided that the election is made before the expiration of 19 months from the priority date. Some designated (or elected) Offices have fixed time limits which expire even later than 20 or 30 months from the priority date. In other Offices an extension of time or grace period, in some cases upon payment of an additional fee, is available.

In addition to these procedural acts, the applicant may also have to comply with other special requirements applicable in certain Offices. **It is the applicant's responsibility** to ensure that the necessary steps to enter the national phase are taken in a timely fashion. Most designated Offices do not issue reminders to applicants in connection with the entry into the national phase.

For detailed information about the procedural acts to be performed to enter the national phase before each designated Office, the applicable time limits and possible extensions of time or grace periods, and any other requirements, see the relevant Chapters of Volume II of the PCT Applicant's Guide. Information about the requirements for filing a demand for international preliminary examination is set out in Chapter IX of Volume I of the PCT Applicant's Guide.

GR and ES became bound by PCT Chapter II on 7 September 1996 and 6 September 1997, respectively, and may, therefore, be elected in a demand or a later election filed on or after 7 September 1996 and 6 September 1997, respectively, regardless of the filing date of the international application. (See second paragraph above.)

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

CONFIRMATION OF PRECAUTIONARY DESIGNATIONS

This notification lists only specific designations made under Rule 4.9(a) in the request. It is important to check that these designations are correct. Errors in designations can be corrected where precautionary designations have been made under Rule 4.9(b). The applicant is hereby reminded that any precautionary designations may be confirmed according to Rule 4.9(c) before the expiration of 15 months from the priority date. If it is not confirmed, it will automatically be regarded as withdrawn by the applicant. There will be no reminder and no invitation. Confirmation of a designation consists of the filing of a notice specifying the designated State concerned (with an indication of the kind of protection or treatment desired) and the payment of the designation and confirmation fees. Confirmation must reach the receiving Office within the 15-month time limit.

REQUIREMENTS REGARDING PRIORITY DOCUMENTS

For applicants who have not yet complied with the requirements regarding priority documents, the following is recalled.

Where the priority of an earlier national, regional or international application is claimed, the applicant must submit a copy of the said earlier application, certified by the authority with which it was filed ("the priority document") to the receiving Office (which will transmit it to the International Bureau) or directly to the International Bureau, before the expiration of 16 months from the priority date, provided that any such priority document may still be submitted to the International Bureau before that date of international publication of the international application, in which case that document will be considered to have been received by the International Bureau on the last day of the 16-month time limit (Rule 17.1(a)).

Where the priority document is issued by the receiving Office, the applicant may, instead of submitting the priority document, request the receiving Office to prepare and transmit the priority document to the International Bureau. Such request must be made before the expiration of the 16-month time limit and may be subjected by the receiving Office to the payment of a fee (Rule 17.1(b)).

If the priority document concerned is not submitted to the International Bureau or if the request to the receiving Office to prepare and transmit the priority document has not been made (and the corresponding fee, if any, paid) within the applicable time limit indicated under the preceding paragraphs, any designated State may disregard the priority claim, provided that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity to furnish the priority document within a time limit which is reasonable under the circumstances.

Where several priorities are claimed, the priority date to be considered for the purposes of computing the 16-month time limit is the filing date of the earliest application whose priority is claimed.

世界知的所有権機関
国際事務局

PCT 特許協力条約に基づいて公開された国際出願

(51) 国際特許分類7 G06F 17/30	A1	(11) 国際公開番号	WO00/43909
		(43) 国際公開日	2000年7月27日(27.07.00)

(21) 国際出願番号 PCT/JP00/00203

(22) 国際出願日 2000年1月18日(18.01.00)

(30) 優先権データ
特願平11/13307 1999年1月21日(21.01.99) JP

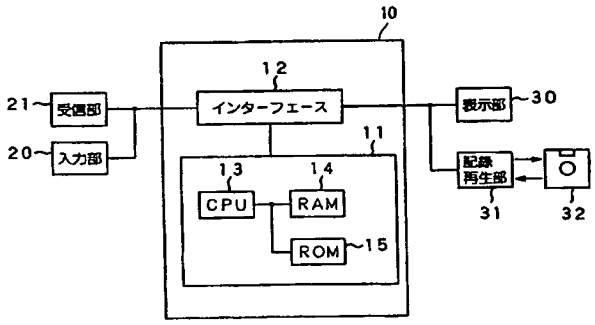
(71) 出願人（米国を除くすべての指定国について）
ソニー株式会社(SONY CORPORATION)[JP/JP]
〒141-0001 東京都品川区北品川6丁目7番35号 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
長尾 確(NAGAO, Katashi)[JP/JP]
〒141-0022 東京都品川区東五反田3丁目14番13号
株式会社 ソニーコンピュータサイエンス研究所内 Tokyo, (JP)
(74) 代理人
小池 晃, 外(KOIKE, Akira et al.)
〒105-0001 東京都港区虎ノ門二丁目6番4号
第11森ビル Tokyo, (JP)

(81) 指定国 JP, US

添付公開書類
国際調査報告書

(54)Title: METHOD AND DEVICE FOR PROCESSING DOCUMENTS AND RECORDING MEDIUM

(54)発明の名称 文書処理方法及び文書処理装置並びに記録媒体



21...RECEPTION UNIT
20...INPUT UNIT
12...INTERFACE
30...DISPLAY UNIT
31...RECORDING/REPRODUCING UNIT

(57) Abstract

A method and device for processing documents each comprised of a plurality of elements and having a tagged internal structure, wherein documents reflecting users' interests are automatically classified, by storing a plurality of documents received by a reception unit in the RAM of a control unit provided in a device body, extracting feature information indicative of features of documents according to the control of the control unit and in conformity with procedures recorded in a ROM, and classifying individual documents by classifying subject in accordance with a level of the relationship between the feature information of documents extracted by a feature information extraction unit in terms of a plurality of classifying subjects constituting a classification model and feature information for each classifying subject.

(57)要約

本発明は、複数のエレメントから構成され、タグ付けされる内部構造を有する文書の処理方法及びその装置であり、受信部で受信した複数の文書を装置本体に設けた制御部のＲＡＭに記憶し、制御部の制御にしたがってＲＯＭに記録された手順にしたがって文書の特徴を表す特徴情報を抽出し、分類モデルを構成する複数の分類項目について特徴情報抽出部で抽出した文書の特徴情報と分類項目毎の特徴情報との関連度に応じて各文書を分類項目に分類することにより、ユーザの興味を反映した文書の自動分類が行われる。

明細書

文書処理方法及び文書処理装置並びに記録媒体

技術分野

本発明は、要素について内部構造を付与された文書を処理する文書処理方法及び文書処理装置並びに文書を処理するプログラムを記録した記録媒体に関し、さらに詳しくは、文書に含まれる語義の関連度に基づいて文書を分類する文書処理方法及び文書処理装置並びに文書に含まれる語義の関連度に基づいて文書を分類する文書処理のプログラムが記録された記録媒体に関する。

背景技術

従来、インターネットにおいて、ウィンドウ形式でハイパーテキスト型情報を提供するアプリケーションサービスとしてWWW（World Wide Web）が用いられている。WWWは、文書の作成、公開又は共有化といった文書処理を実行し、新しいスタイルの文書の在り方を示したシステムである。文書の実際上の利用の観点からは、文書の内容に基づいた文書の分類や要約といった、WWWを越える高度な文書処理が求められている。このような高度な文書処理には、文書の内容の機械的な処理が不可欠である。

文書の内容の機械的な処理は、以下のような理由から依然として困難である。第１に、ハイパーテキストを記述する言語であるＨＴ

ＭＬ（Hyper Text Markup Language）は、文書の表現については規定するが、文書の内容についてはほとんど規定しない。第２に、文書間に構成されたハイパーテキストのネットワークは、文書の読者にとって文書の内容を理解するために必ずしも利用しやすいものではない。第３に、一般に文章の著作者は読者の便宜を念頭に置かずに著作するが、文書の読者の便宜が著作者の便宜と調整されることはない。

ＷＷＷは、新しい文書の在り方を示したシステムであるが、文書を機械的に処理しないために、高度な文書処理を行うことができない。ＷＷＷにおいて、高度な文書処理を実行するためには、文書を機械的に処理することが必要となる。

そこで、文書の機械的な処理を可能とするため、文書の機械的な処理を支援するシステムが自然言語研究の成果に基づいて開発されている。自然言語研究による文書処理の最初のステップとして、文書の著作者等による文書の内部構造についての属性情報であるタグの付与を前提とした文書に付与されたタグを利用する機械的な文書処理が提案されている。

ところで、コンピュータの普及やネットワーク化の進展に伴い、文章処理や、文書の内容に依存した索引などで、テキスト文書の作成、ラベル付け、変更などを行う文書処理の高機能化が求められている。このような高機能な文書処理を実現するためには、文書内における各語義の関連度に基づいた文書処理が必要となる。

発明の開示

本発明は、上述の実情に鑑みて提案されるものであって、文書内における語義の関連度に基づいた文書処理を行うような文書処理方法及び文書処理装置並びに文書内における関連度に基づいた文書処理のプログラムを記録した記録媒体を提供することを目的とする。

このような目的を達成するために提案される本発明は、複数の要素から構成される内部構造を有する文書を処理する文書処理方法であり、文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出する工程と、分類モデルを構成する複数の分類項目について特徴情報抽出工程で抽出した文書の特徴情報と分類項目毎の特徴情報との関連度に応じて各文書を分類項目に分類する文書分類工程とを有する。

また、本発明は、複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と、重み付与工程で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、他の語義に付与された重みの差分を一の語義に付与された重みの差分で除した商を一の語義と他の語義の関連度として演算する関連度演算工程とを有する。

さらに、本発明は、複数の要素から構成される内部構造を有する文書を処理する文書処理装置であり、文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出部と、分類モデルを構成する複数の分類項目について特徴情報抽出部で抽出した文書の特徴情報と分類項目毎の特徴情報との関連度に応じて各文書を分類項目に分類する文書分類部とを有する。

さらにまた、本発明は、複数の要素から構成される内部構造を有

する文書を処理する文書処理装置において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与部と、重み付与部で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし他の語義に付与された重みの差分を一の語義に付与された重みの差分で除した商を一の語義と他の語義の関連度として演算する関連度演算部とを有する。

さらにまた、本発明は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体であり、この記録媒体に記録されるプログラムは、文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、分類モデルを構成する複数の分類項目について特徴情報抽出処理で抽出した文書の特徴情報と分類項目毎の特徴情報との関連度に応じて各文書を分類項目に分類する文書分類処理とを行う。

さらにまた、本発明は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体であり、この記録媒体に記録されるプログラムは、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と、重み付与処理で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし他の語義に付与された重みの差分を一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理とを行う。

本発明の更に他の目的、本発明によって得られる具体的な利点は、以下に説明される実施例の説明から一層明らかにされるであろう。

図面の簡単な説明

図１に本発明が適用された文書処理装置を示すブロック図である。

図２は、文書のタグ付けによる内部構成を示すツリー図である。

図３は、文書のタグ付けによる内部構成を表示したウィンドウを示す平面図である。

図４は、本発明に係る文書処理装置の動作を示すフローチャートである。

図５は、文書の自動分類を行うＧＵＩを示す平面図である。

図６は、文書を自動分類するフローチャートである。

図７は、文書の特徴を発見してインデックスを作成するフローチャートである。

図８は、活性拡散を示すフローチャートである。

図９は、活性拡散の処理を説明する図である。

図１０は、活性拡散のリンク処理のフローチャートである。

図１１は、語義の関連度の計算のフローチャートである。

図１２は、語義の関連度の表を示す図である。

図１３は、分類モデルの表を示す図である。

図１４は、関連度による文書分類のフローチャートである。

発明を実施するための最良の形態

以下、本発明に係る文書処理方法及び文書処理装置並びに記録媒体を図面を参照して具体的に説明する。

本発明に係る文書処理装置は、図１に示すように、制御部１１とインターフェース１２を備える本体１０と、ユーザからの入力を受

け付けて本体１０に送る入力部２０と、外部からの信号を受信して本体１０に送る受信部２１と、本体１０からの出力を表示する表示部３０と、記録媒体３２に対して情報を記録し記録媒体３２に記録された情報の再生を行う記録再生部３１とを有している。

本体１０は、制御部１１とインターフェース１２を有し、この文書処理装置の主要な部分を構成している。制御部１１は、この文書処理装置における処理を集中して実行するＣＰＵ１３と、揮発性のメモリであるＲＡＭ１４と、不揮発性のメモリであるＲＯＭ１５とを有している。ＣＰＵ１３は、ＲＯＭ１５に記録された手順にしたがって必要な場合にはデータを一時的にＲＡＭ１４に格納して、プログラムを実行するための制御を行う。インターフェース１２には、入力部２０、受信部２１及び表示部３０が接続される。インターフェース１２は、制御部１１からの制御の下に、入力部２０及び受信部２１からのデータの入力、表示部３０へのデータの送信について、データを送信するタイミングを調整し、データの形式を変換する。

入力部２０は、この文書処理装置に対するユーザの入力を受け付ける部分である。この入力部２０は、例えばキーボードやマウスにより構成される。ユーザは、この入力部２０を用い、キーボードによりキーワードを入力し、マウスにより表示部３０に表示されている文書のエレメントを選択して入力することができる。ここで、エレメントとは、文書を構成する要素であって、例えば文書、文及び語が含まれる。

受信部２１は、この文書処理装置に外部から例えば通信回線を介して送信される信号を受信する部分である。この受信部２１は、例えば電子文書である複数の文書を受信する。受信部２１は、受信し

たデータを本体１０に送る。

出力部３０は、この文書処理装置からの出力結果を表示するものである。この出力部３０は、陰極線管（cathode ray tube;CRT）や液晶表示装置（liquid crystal display;LCD）から構成され、単数又は複数のウィンドウを表示し、このウィンドウ上に文字、図形等を表示する。

記録再生部３１は、この文書処理装置の制御部１１により制御されて、フロッピーディスクのような記録媒体３２に対して情報の記録再生を行う。記録媒体３２には、例えば文書の語義に基づいて関連度を求め、この関連度に基づいて文書処理を実行するようなプログラムが記録されている。この記録媒体３２の詳細については、さらに後述する。

続いて、本発明において取り扱われる文書について説明する。この文書は、ツリー状のタグ付けによる内部構造を有している。本発明では、図２に示すように、タグ付けによる内部構造、文書、文、語彙エレメント等の各エレメント、通常リンク、参照・被参照リンク等がタグとしてあらかじめ文書に付与されている。図２において、図中において、白丸“○”は文書の要素、すなわちエレメントであり、最下位の白丸は文書における最小レベルの語に対応する、語彙エレメントである。また、実線は語、句、節、文等の文書の構造を示す通常リンク（normal link ）である。破線は参照・被参照による係り受け関係を示す参照リンク（reference link）である。文書のタグ付けによる内部構造は、上位から下位への順序で、文書（document）、文書の下位であり段落の上位であるオプションのサブディビジョン（subdivision ）、オプションの段落（paragraph）、文

（sentence ）、文の下位であるサブセンテンシャルセグメント（sub-sentential segment ）、・・・、最下位の語彙エレメントのような階層構造である。

ここで、文書のタグ付けによる内部構造として、多言語間に共通な意味的・語用論的タグを文書に付与することにより、文書の機械的な内容理解を可能にするようなタグ付けを採用している。タグとは、データに対してその属性を表すために付加される属性情報である。

文書のタグ付けによる内部構造は、ＨＴＭＬ（Hyper Text Markup Language）と同様にＸＭＬ（Extended Markup Language）の形式のタグである。すなわち、タグは、係り受け、例えば代名詞の指示対象、多義語の意味のように統語（syntactic）・意味（semantic）等の情報を含んでいる。

文章のタグ付けによる内部構造の一例を次に示すが、文章へのタグ付けはこの方法に限られない。

例えば、“Time flies like an arrow.”という文については、

＜文＞＜名詞句　語義＝“time0”＞time＜／名詞句＞
＜動詞句＞＜動詞　語義＝“fly1”＞flies＜／動詞＞
＜形容動詞句＞＜形容動詞　語義＝like0＞like＜／形容動詞＞
＜名詞句＞an＜名詞　語義＝“arrow0”＞arrow＜／名詞＞＜／名詞句＞＜／形容動詞句＞＜／動詞句＞.＜／文＞

というようにタグ付けすることができる。ここで＜文＞、＜名詞＞、＜名詞句＞、＜動詞＞、＜動詞句＞、＜形容動詞＞、＜形容動詞句＞は、それぞれ一文、名詞、名詞句、動詞、動詞句、前置詞句、後置詞句を含む形容詞／形容詞句、形容詞句／形容動詞句のような文

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTICE INFORMING THE APPLICANT OF THE COMMUNICATION OF THE INTERNATIONAL APPLICATION TO THE DESIGNATED OFFICES

(PCT Rule 47.1(c), first sentence)

To:
KOIKE, Akira
No.11 Mori Building
6-4, Toranomon 2-chome
Minato-ku, Tokyo 105-0001
JAPON

Date of mailing (day/month/year)	
27 July 2000 (27.07.00)	
Applicant's or agent's file reference SK00PCT7	**IMPORTANT NOTICE**
International application No. PCT/JP00/00203	**International filing date** (day/month/year) 18 January 2000 (18.01.00)
Priority date (day/month/year) 21 January 1999 (21.01.99)	
Applicant SONY CORPORATION et al	

1. Notice is hereby given that the International Bureau has communicated, as provided in Article 20, the international application to the following designated Offices on the date indicated above as the date of mailing of this Notice:
   JP,US

   In accordance with Rule 47.1(c), third sentence, those Offices will accept the present Notice as conclusive evidence that the communication of the international application has duly taken place on the date of mailing indicated above and no copy of the international application is required to be furnished by the applicant to the designated Office(s).

2. The following designated Offices have waived the requirement for such a communication at this time:
   None

   The communication will be made to those Offices only upon their request. Furthermore, those Offices do not require the applicant to furnish a copy of the international application (Rule 49.1(a-bis)).

3. Enclosed with this Notice is a copy of the international application as published by the International Bureau on
   27 July 2000 (27.07.00) under No. WO 00/43909

REMINDER REGARDING CHAPTER II (Article 31(2)(a) and Rule 54.2)

If the applicant wishes to postpone entry into the national phase until 30 months (or later in some Offices) from the priority date, **a demand for international preliminary examination** must be filed with the competent International Preliminary Examining Authority before the expiration of 19 months from the priority date.

It is the applicant's sole responsibility to monitor the 19-month time limit.

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

REMINDER REGARDING ENTRY INTO THE NATIONAL PHASE (Article 22 or 39(1))

If the applicant wishes to proceed with the international application in the **national phase**, he must, within 20 months or 30 months, or later in some Offices, perform the acts referred to therein before each designated or elected Office.

For further important information on the time limits and acts to be performed for entering the national phase, see the Annex to Form PCT/IB/301 (Notification of Receipt of Record Copy) and Volume II of the PCT Applicant's Guide.

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 740.14.35

Authorized officer
J. Zahra

Telephone No. (41-22) 338.83.38

世界知的所有権機関
国際事務局

PCT 特許協[redacted]条約に基づいて公開された国[redacted]出願

(51) 国際特許分類7
G06F 17/30

A1

(11) 国際公開番号 WO00/43909

(43) 国際公開日 2000年7月27日(27.07.00)

(21) 国際出願番号 PCT/JP00/00203

(22) 国際出願日 2000年1月18日(18.01.00)

(30) 優先権データ
特願平11/13307 1999年1月21日(21.01.99) JP

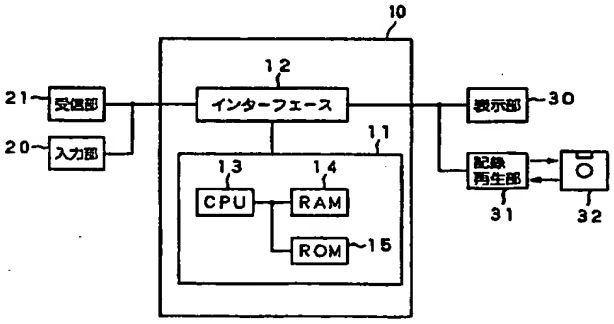
(71) 出願人（米国を除くすべての指定国について）
ソニー株式会社(SONY CORPORATION)[JP/JP]
〒141-0001 東京都品川区北品川6丁目7番35号 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
長尾 確(NAGAO, Katashi)[JP/JP]
〒141-0022 東京都品川区東五反田3丁目14番13号
株式会社 ソニーコンピュータサイエンス研究所内 Tokyo, (JP)
(74) 代理人
小池 晃, 外(KOIKE, Akira et al.)
〒105-0001 東京都港区虎ノ門二丁目6番4号
第11森ビル Tokyo, (JP)

(81) 指定国 JP, US

添付公開書類
国際調査報告書

(54)Title: METHOD AND DEVICE FOR PROCESSING DOCUMENTS AND RECORDING MEDIUM

(54)発明の名称 文書処理方法及び文書処理装置並びに記録媒体



21...RECEPTION UNIT
20...INPUT UNIT
12...INTERFACE
30...DISPLAY UNIT
31...RECORDING/REPRODUCING UNIT

(57) Abstract

A method and device for processing documents each comprised of a plurality of elements and having a tagged internal structure, wherein documents reflecting users' interests are automatically classified, by storing a plurality of documents received by a reception unit in the RAM of a control unit provided in a device body, extracting feature information indicative of features of documents according to the control of the control unit and in conformity with procedures recorded in a ROM, and classifying individual documents by classifying subject in accordance with a level of the relationship between the feature information of documents extracted by a feature information extraction unit in terms of a plurality of classifying subjects constituting a classification model and feature information for each classifying subject.

(57)要約

本発明は、複数のエレメントから構成され、タグ付けされる内部構造を有する文書の処理方法及びその装置であり、受信部で受信した複数の文書を装置本体に設けた制御部のＲＡＭに記憶し、制御部の制御にしたがってＲＯＭに記録された手順にしたがって文書の特徴を表す特徴情報を抽出し、分類モデルを構成する複数の分類項目について特徴情報抽出部で抽出した文書の特徴情報と分類項目毎の特徴情報との関連度に応じて各文書を分類項目に分類することにより、ユーザの興味を反映した文書の自動分類が行われる。